

# **Complete mitochondrial DNA genome variation in Peninsular Malaysia**

Ken Khong Eng

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

School of Biology  
Faculty of Biological Sciences  
University of Leeds

January 2014

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

© 2014 The University of Leeds and Ken Khong Eng

The right of Ken Khong Eng to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

## **Acknowledgements**

I would like to thank Martin Richards and Stephen Oppenheimer for the supervision, help and guidance they have provided throughout this research. I would like to thank Mokhtar Saidin and Stephen Chia for their advice and support, especially for encouraging me to apply for the fellowship of Academic Staff Training Scheme, Universiti Sains Malaysia. I would like to acknowledge Zafarina Zainuddin for kindly providing the valuable Malay samples.

I am very grateful to the help I received in the lab from Pedro Soares, Maria Pala, Martin Carr, Marta Costa, Joana Pereira and Verónica Fernandes; enormous thanks to their patience in showing me how to run the phylogenetic software in this study. I also learn a lot from our lab meetings, I am definitely going to miss it.

Last but not least, I would like to thank my family and friends for your endless love, understanding and massive support, also for putting up with me during the more stressful time!

Thank you!

## Abstract

The peopling of Southeast Asia has been vigorously debated over the past few decades by archaeologists, linguists and anthropologists, as well as evolutionary and population geneticists. Several ethnic minorities in the region, the *Orang Asli* groups (the Semang, Senoi and Aboriginal Malays) from Peninsular Malaysia, are widely thought of as “relicts” of human diversity in the ancient Sunda continent. However, mitochondrial DNA (mtDNA) analysis of these groups has hitherto been restricted to a small number of populations and largely based on the mtDNA control region hypervariable segment I (HVS-I), supplemented by a very small number of whole-mtDNA genomes. In this study, I have both expanded the number of populations examined and analysed 226 lineages at the level of whole-mtDNA genomes from both *Orang Asli* and modern Malay populations, covering most of the extant mtDNA diversity in Peninsular Malaysia, in the context of Southeast Asian variation more generally, including a total of 2206 complete mtDNA sequences in the phylogeographic analysis. This has confirmed that the *Orang Asli* populations indeed experienced high genetic drift, likely due to their extremely small group sizes and population subdivision. All three *Orang Asli* groups have local roots that trace back to ~50 ka, and all have been affected to a greater or lesser extent by subsequent migrations to Peninsular Malaysia. The Semang and Senoi show much less haplogroup diversity than the Aboriginal Malays, although the latter have some indigenous ancestry that is as deep as that of the Semang and Senoi in Peninsular Malaysia. However, this drift, and the loss of lineages that it has entailed, is compensated for by the retention of many related ancient lineages in the extant modern Malay, who therefore provide a more comprehensive view of ancient Malay Peninsula, and more generally ancient Sunda, mtDNA diversity. Indeed, contrary to the model that posits a recent ancestry for Malay in Island Southeast Asia (ISEA), a majority of their maternal lineages appear to have had a local ancestry within Mainland Southeast Asia (MSEA) and the Malay Peninsula. Combining the *Orang Asli* and Malay data indicates a very deep ancestry for multiple indigenous maternal lineages that date back locally (or regionally) to the late Pleistocene. Many can be traced to the original inhabitants of Southeast Asia, who colonised the Sunda region from South Asia ~50–60 ka. It appears that the spread of the so-called “Coastal Neolithic” foraging groups (who may have engaged in horticulture, but were largely pre-rice agriculture) may have provided the main contribution to the north–south lineage expansions,



and the spread of Austro-Asiatic languages to the *Orang Asli* and to the Nicobars may be connected to some of these dispersals. Apart from preserving these ancient lineages, many of which have been lost by drift in the relict populations, the modern Malay also preserve complex maternal influences from further afield at various times stretching back to the Last Glacial Maximum, from ISEA (as far east as the New Guinea region), to a lesser extent from East Asia, and to an even lesser extent South Asia. Climatic change and sea-level rises were likely the most important driving force behind the demographic history of Southeast Asia, mainland as well as insular, as shown by a sharp signal of early Holocene population crash and subsequent re-expansion in both the modern Malay and the *Orang Asli*. Although there is substantial lineage sharing between modern Malay and their close Sunda neighbours in Sumatra, ISEA lineages amount to little more than a quarter of the maternal variation of Malay, and even if there was a major migration to the Peninsula in the Late Holocene, the majority of their maternal ancestry seems to lie within the bounds of the Sunda continent.

## Table of Content

<b>Acknowledgements.....</b>	<b>iii</b>
<b>Abstract .....</b>	<b>iv</b>
<b>Table of Content.....</b>	<b>vi</b>
<b>List of Figures.....</b>	<b>xii</b>
<b>List of Tables .....</b>	<b>xxiii</b>
<b>List of Abbreviations.....</b>	<b>xxiv</b>
<b>1 Introduction.....</b>	<b>1</b>
1.1 Mitochondria .....	1
1.1.1 Mutation rates .....	4
1.1.2 Studies using human mtDNA .....	4
1.2 The Human mtDNA Phylogenetic Tree .....	6
1.2.1 mtDNA tree nomenclature .....	7
1.2.2 The mtDNA molecular clock.....	9
1.3 Y chromosome haplogroup phylogeny .....	10
1.4 Autosomal Markers.....	13
1.5 The Origin of Anatomically Modern Humans (AMH).....	15
1.6 Out of Africa .....	16
1.7 The Southern Route .....	18
1.8 Southeast Asia .....	20
1.8.1 The flooding of Sundaland .....	20
1.8.2 First settlement of ISEA by modern humans.....	23
1.8.3 The People of SEA.....	24
1.8.4 The “Out of Taiwan model” and the “Farming/Language Dispersal Hypothesis” .....	27
1.9 <i>Orang Asli</i> in Peninsular Malaysia .....	33
1.9.1 The “Negrito Hypothesis” .....	38
1.10 Modern Malay in Peninsular Malaysia .....	40
1.11 Previous Phylogeographic Analysis .....	40
1.12 Objectives and Hypotheses .....	47

<b>2</b>	<b>Material and Methods .....</b>	<b>49</b>
2.1	Samples .....	49
2.1.1	Participants in this study.....	49
2.1.2	Comparative published mtDNA complete sequences.....	51
2.2	Phenol-chloroform DNA extraction .....	52
2.3	Whole-genome amplification .....	52
2.4	Polymerase Chain Reaction (PCR) Amplification and Sequencing.....	53
2.5	Gel electrophoresis .....	57
2.6	DNA purification and sequencing .....	57
2.7	Data Manipulation .....	58
2.7.1	Variants scoring .....	58
2.7.2	Error detection .....	59
2.7.3	v2nall.....	59
2.7.4	Geneious .....	59
2.7.5	Alignments by Clustal algorithm.....	60
2.7.6	mtDNA-GeneSyn tool.....	60
2.7.7	fm2net_gui.....	60
2.8	Phylogeographic Analysis.....	61
2.8.1	Network 4.6 .....	62
2.8.2	Phylogenetic trees .....	63
2.9	Coalescence time estimation .....	65
2.9.1	The rho ( $\rho$ ) statistic .....	65
2.9.2	Maximum Likelihood (ML) .....	65
2.9.3	Bayesian evolutionary analysis by sampling trees (BEAST) .....	67
<b>3</b>	<b>Results and Discussion: Control Region and Haplogroup M .....</b>	<b>70</b>
3.1	Control-region variation.....	70
3.2	Haplogroup M .....	81
3.3	Haplogroup M21.....	82
3.4	Haplogroup M22.....	86
3.5	Haplogroup M7 .....	88
3.5.1	Haplogroup M7a .....	88
3.5.2	Haplogroup M7c'e'f.....	88
3.5.3	Haplogroup M7b'd'g .....	92
3.6	Haplogroup M9 .....	95

3.6.1	Haplogroup M9a'b.....	96
3.6.2	Haplogroup E1.....	99
3.6.3	Haplogroup E2.....	103
3.7	Haplogroup M17.....	105
3.8	Haplogroup M12'G.....	107
3.8.1	Haplogroup G1 .....	108
3.8.2	Haplogroup G2 .....	109
3.8.3	Haplogroups G3 and G4.....	110
3.8.4	Haplogroup M12.....	110
3.9	Haplogroup M29'Q.....	112
3.10	Haplogroup M8 .....	114
3.11	Haplogroup M4'67 .....	117
3.12	Haplogroup M5 .....	118
3.13	Haplogroup M24'41 .....	119
3.14	Haplogroup M13'46'61 .....	120
3.15	Haplogroup M2 .....	123
3.16	Haplogroup M32'56 .....	125
3.17	Haplogroup M47.....	126
3.18	Haplogroup M26.....	126
3.19	Haplogroup M1'20'51 .....	128
3.20	Haplogroup M50.....	131
3.21	Haplogroup M71.....	132
3.22	Haplogroup M72.....	133
3.23	Haplogroup M42'74 .....	134
3.24	Haplogroup M73'79 .....	136
3.25	Haplogroup M77.....	138
3.26	Haplogroup D.....	138
3.26.1	Haplogroup D4.....	140
3.26.2	Haplogroup D5.....	140
3.26.3	Haplogroup D6.....	142
3.27	Novel M* Haplogroups.....	144
<b>4</b>	<b>Results and Discussion: Haplogroup N.....</b>	<b>145</b>
4.1	Haplogroup N9 .....	145
4.1.1	Haplogroup N9a.....	146

4.1.2	Haplogroup N9b.....	149
4.1.3	Haplogroup Y .....	149
4.2	Haplogroup N21 .....	151
4.3	Haplogroup N22 .....	152
4.4	Haplogroup N8 .....	154
4.5	Haplogroup N10 .....	155
4.6	Haplogroup N11 .....	156
4.7	Haplogroup A.....	158
<b>5</b>	<b>Results and Discussion: Haplogroup R.....</b>	<b>162</b>
5.1	Haplogroup B .....	162
5.1.1	Haplogroup B4+C16261T .....	163
5.1.2	Haplogroups B4b'd'e'j and B4f .....	170
5.1.3	Haplogroup B4c .....	172
5.2	Haplogroup B5 .....	175
5.2.1	Haplogroup B5a.....	176
5.2.2	Haplogroup B5b.....	178
5.3	Haplogroup R11'B6.....	183
5.4	Haplogroup R12'21 .....	184
5.5	Haplogroup R22 .....	185
5.6	Haplogroup R9 .....	187
5.6.1	Haplogroup R9b.....	188
5.6.2	Haplogroup R9c1 .....	190
5.6.3	Haplogroup F1 .....	191
5.6.4	Haplogroup F2 .....	197
5.6.5	Haplogroup F3 .....	197
5.6.6	Haplogroup F4 .....	199
5.7	Haplogroup P.....	199
5.8	Haplogroup R6 .....	202
5.9	Haplogroup R7 .....	203
5.10	Haplogroup R23 .....	204
5.11	Haplogroup R30 .....	205
5.12	Haplogroup U .....	205

<b>6</b>	<b>Bayesian Skyline Plot (BSP).....</b>	<b>207</b>
<b>7</b>	<b>Discussion and Conclusions .....</b>	<b>211</b>
7.1	Semang and Senoi.....	213
7.1.1	The “Negrito Hypothesis” .....	217
7.2	Aboriginal Malays (aka ‘Proto-Malays’).....	220
7.3	Modern Malay (aka ‘Deutero-Malay’) .....	222
7.3.1	MSEA/Sunda haplogroups in the modern Malay .....	224
7.3.2	ISEA haplogroups in the modern Malay .....	228
7.3.3	East Asian haplogroups in the modern Malay .....	229
7.3.4	South Asian haplogroups in the modern Malay .....	230
7.4	Conclusions .....	231
	<b>References.....</b>	<b>235</b>
	<b>Appendices .....</b>	<b>259</b>
	Appendix A.....	259
	Appendix B .....	262
	Appendix C .....	264
	Appendix D.....	267
	Appendix E – Haplogroups not found in Malaysia.....	280
	Haplogroup M7a.....	280
	Haplogroup M9a1a1c.....	281
	Haplogroup G2 .....	281
	Haplogroups G3 and G4 .....	283
	Haplogroup M8a.....	285
	Haplogroups C1 and C4 .....	286
	Haplogroup M10.....	288
	Haplogroup M60.....	289
	Haplogroup M76.....	291
	Haplogroup D4 .....	291
	Haplogroup D5a .....	299
	Haplogroup N9a2’4’5 .....	300
	Haplogroup N9b .....	301
	Haplogroup A5a .....	302
	Haplogroups B4d and B4f.....	303
	Haplogroup B4c1.....	304

Haplogroups F1b, F1d and F1e .....	304
Haplogroup F2.....	307
Haplogroup P.....	307
Haplogroup R30 .....	308

## List of Figures

Figure 1.1 Schematic diagram of the human mitochondrial genome. The genome encodes 22 transfer RNAs (indicated by single letter abbreviation) between the coding genes, two ribosomal RNAs (12S and 16S), and 13 essential genes that encode subunits of the oxidative phosphorylation enzyme complexes. The D-loop region includes heavy and light-strand promoters (HSP and LSP), and the origin of L-strand replication (O <sub>L</sub> ).....	2
Figure 1.2 Diagram of the mammalian mtDNA control region (Modified from Chinnery, 2006). The control region consists of three hypervariable segments (I, II and III) separated by three conserved sequence blocks (CSBs). LSP and HSP are the promoters for the L and H- strand respectively. ....	2
Figure 1.3 Structure of a rooted tree. A, B, C, D, E, and F are external nodes. G, H, I, J, and K are internal nodes, with K as root node. In a rooted tree, the arrow indicates the direction of evolution. Figure adapted from Vandamme (2009). ....	5
Figure 1.4 The simplified view of the global human mtDNA phylogenetic tree Build 15 version (30 Sep 2012). The root of the tree (mt-MRCA) represents the maternal most recent common ancestor of all humans. Haplogroups L0 to L6 are lineages specific to Africa, except that L3 diverged into M and N from which all the remaining diversity is derived. For more details, see: <a href="http://www.phylotree.org">www.phylotree.org</a> , van Oven and Kayser (2009).....	8
Figure 1.5 Map showing the inferred movements of mtDNA haplogroups in Africa and Eurasia between 60 to 30 ka. The figure indicates the African exodus on the Southern route ~ 60 ka, taking the coastal path along the Indian Ocean into Southeast Asia and Australasia. Figure modified from Family Tree DNA (2006), Soares et al. (2009) and Beyin (2011). ....	17
Figure 1.6 Sundaland in the Late Pleistocene period. Areas in yellow were drowned when the sea-level rose; brown areas indicate the present day countries. Figure adapted from Oppenheimer (1998).....	22
Figure 1.7 The area highlighted shows the extent of the Austronesian migrations. Figure adapted from Quirino (2010).....	25
Figure 1.8 Austronesian languages with corresponding geographical location. Figure after Blust (1995). ....	29
Figure 1.9 Map of Peninsular Malaysia showing the locations of Orang Asli groups sampled. Map adapted from Oppenheimer (1998). ....	33
Figure 2.1 Locations of the samples included in this study. Key: Kedah (KDH), Kelantan (KEL), Perak (PRK), Negeri Sembilan (NS) and Johor (JHR) in Peninsular Malaysia; Kota Kinabalu (KK), Brunei (BRU), Palangkaraya (PRY) and Banjarmasin (BAN) in Borneo; Padang (PAD), Palembang (PLB) and Bangka (BGK) in Sumatra; Tengger (TGR) in Jawa Timur; Bali (BAL), Mataram (MTR), Waingapu (WAI), Alor (ALO) in Nusa Tenggara Timur; Palu (PAL), Ujung Pandang (UJP) and Manado (MND) in Sulawesi; Thailand (THA); Vietnam (VNM); Philippines (FIL); Aboriginal Taiwanese (TAI). ....	50
Figure 2.2 fm2net_gui interface. ....	61



Figure 2.3 The colour codes for samples divided according to regional locations. ....	64
Figure 2.4 A screenshot of the baseml.ctf file opened in Notepad.....	66
Figure 2.5 Settings for Bayesian Skyline Plot in Tracer software. ....	69
Figure 3.1 The HVS-I Reduced-median network of haplogroup M for 34 samples. (Label boxes in light grey indicate Semang, and light blue for Senoi).....	73
Figure 3.2 The HVS-I Reduced-median network for haplogroup N and R for 48 samples. One Semelai of Aboriginal Malay sample (ORA; Hill et al., 2006) from haplogroup F1a1a with a transition at np 16304 was included in the network to define the R root. (Label boxes in light blue indicate Senoi, light grey for Semang, and light yellow for Aboriginal Malay) .....	73
Figure 3.3 The HVS-I reduced-median network of all haplogroups found in K.C. Ang's 18 Orang Asli subgroups. ....	75
Figure 3.4 The HVS-I corrected reduced-median network of all haplogroup found in K.C. Ang's 18 Orang Asli subgroups.....	75
Figure 3.5 The HVS-I reduced-median network for haplogroup M of K.C. Ang's 37 samples. (Label boxes in light blue indicate Senoi, light grey for Semang, and light yellow for Aboriginal Malay) .....	75
Figure 3.6 The HVS-I reduced-median network for haplogroups N and R of K.C. Ang's 56 samples. The roots N and R are indicated in the network.....	76
Figure 3.7 Schematic diagram of haplogroup M's major subclades present in Southeast Asia. ....	81
Figure 3.8 The tree of haplogroup M21a. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BAT – Semang Batek, FIL – Philippines, JAH – Semang Jahai, KEN – Semang Kensiu, KIN – Semang Kintak, LAN – Semang Lanoh, MEN – Semang Mendriq, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SBO – South Borneo, SML – Aboriginal Malay Semelai, SUL – Sulawesi, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, THA – Thailand, VIE – Vietnam).....	82
Figure 3.9 The tree of haplogroup M21c'd. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FBT – Philippines Batak, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SML – Aboriginal Malay Semelai, THA – Thailand, VIE – Vietnam, TMK – Thailand Moken) .....	84
Figure 3.10 HVS-I network of M21*. M21b has been reassigned to M13b. Figure adapted from Hill (2005). ....	85
Figure 3.11 The tree of haplogroup M22. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, VIE – Vietnam).....	87
Figure 3.12 HVS-I network of M22. Figure adapted from Hill (2005). ....	87
Figure 3.13 Schematic diagram of haplogroup M7 and its major subclades distribution. (EA – East Asia, JAP – Japan, SAS – South Asia, SEA – Southeast Asia, TEM – Aboriginal Malay Temuan) .....	88

Figure 3.14 The tree of haplogroup M7c'e'f excluding M7c3c. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan, MGL – Inner Mongolia, China, TEM – Aboriginal Malay Temuan, XIN – Xinjiang, China) .....	89
Figure 3.15 The tree of haplogroup M7c3c. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. Mutation at np 310 in the GU733- sequences (Gunnarsdóttir et al., 2011a) was removed from the tree for posing incorrect evolutionary pathways. (BID – Bidayuh Sarawak, FIL – Philippines, MIC – Micronesia, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUM – Sumatra, TAI – Taiwan).....	90
Figure 3.16 HVS-I network of M7c3c. Figure adapted from Hill (2005). .....	91
Figure 3.17 The tree of haplogroup M7b'd'g (excluding M7b1'2'4-8+16189). Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. The Philippines sequences marked by “*” are excluded from age estimations. (CHI – China, FIL – Philippines, IND – India, SEM – Southeast Peninsular Malay, TAI – Taiwan, TEM – Aboriginal Malay Temuan, THA – Thailand).....	93
Figure 3.18 The tree of haplogroup M7b1'2'4-8. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NWM – Northwest Peninsular Malay, SWM – Southwest Peninsular Malay, XIN – Xinjiang, China) .....	94
Figure 3.19 Weighted HVS-I network of M7b*, M7b1 and M7b3 types with the root indicated. Figure adapted from Hill (2005).....	94
Figure 3.20 Schematic diagram of haplogroup M9 and its major subclades distribution. (CAS – Central Asia, EA – East Asia, FIL – Philippines, ISEA – Island Southeast Asia, MSEA – Mainland Southeast Asia, PNG – Papua New Guinea, SAS – South Asia, SC – South China, SEA – Southeast Asia, TAI – Taiwan).....	96
Figure 3.21 The tree of haplogroup M9a'b, excluding M9a1. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NBO – North Borneo, NWM – Northwest Peninsular Malay, TAI – Taiwan, VIE – Vietnam, XIN – Xinjiang) .....	97
Figure 3.22 The tree of haplogroup M9a1, excluding M9a1a1c. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BGL – Bangladesh, CHI – China, IND – India, JAP – Japan, JAV – Java, Indonesia, MGL – Inner Mongolia, China, MYA – Myanmar, TIB – Tibet, XIN – Xinjiang) .....	97
Figure 3.23 The phylogeny of haplogroup E1 excluding E1a1a and E1a2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BID – Bidayuh Sarawak, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NBO – North Borneo, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SBO – South Borneo, SUL – Sulawesi, PNG – Papua New Guinea, TAI – Taiwan).....	100
Figure 3.24 The phylogeny of haplogroup E1a1a. Sequence marked by “*” are excluded from age estimations. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, MAM – Philippines Mamanwa, NBO – North Borneo, NEM – Northeast Peninsular Malay,	

NWM – Northwest Peninsular Malay, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, PNG – Papua New Guinea, TAI – Taiwan).....	101
Figure 3.25 The phylogeny of haplogroup E1a2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, NBO – North Borneo, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUL – Sulawesi, PNG – Papua New Guinea).....	101
Figure 3.26 HVS-I network of E1* and E1b. Figure adapted from Hill (2005).....	103
Figure 3.27 HVS-I network of E1a1a. Figure adapted from Hill (2005). ....	103
Figure 3.28 The tree of haplogroup E2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. GU733721 is excluded from age estimations. (FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, MOL – Moluccas, Indonesia, NBO – North Borneo, PNG – Papua New Guinea, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUM – Sumatra).....	104
Figure 3.29 The tree of haplogroup M17. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CAM – Cambodia, FIL – Philippines, INA – Indonesia, JAV – Java, Indonesia, KEN – Semang Kensi, NEM – Northeast Peninsular Malay, SBO – South Borneo, SUL – Sulawesi, THA – Thailand, VIE – Vietnam).....	106
Figure 3.30 Schematic diagram of haplogroup M12'G and its major subclades distribution. (CAS – Central Asia, EA – East Asia, ISEA – Island Southeast Asia, JAP – Japan, MSEA – Mainland Southeast Asia, NAS – North Asia, NEM – Northeast Peninsular Malay, SAS – South Asia, SC – South China, SEA – Southeast Asia, SLT – Aboriginal Malay Seletar).....	108
Figure 3.31 The tree of haplogroup G1. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, JAP – Japan, MGL – Inner Mongolia, China, SLT – Aboriginal Malay Seletar) .....	109
Figure 3.32 The tree of haplogroup M12. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, IND – India, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SUM – Sumatra, THA – Thailand, VIE – Vietnam) .....	111
Figure 3.33 The tree of haplogroup M29'Q. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BOU – Bougainville, COO – Cook Island, FIL – Philippines, NWM – Northwest Peninsular Malay, PNG – Papua New Guinea, SEM – Southeast Peninsular Malay, SMO – Samoa, VAN – Vanuatu) .....	113
Figure 3.34 HVS-I network of Q1. Figure adapted from Hill (2005). ....	114
Figure 3.35 Schematic diagram of haplogroup M8 and the distribution of its major subclades. (AME – America, CAS – Central Asia, CHI – China, EA – East Asia, FIL – Philippines, JAP – Japan, MSEA – Mainland Southeast Asia, NAS – North Asia, SC – South China) .....	115
Figure 3.36 The tree of haplogroups C5 and C7. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NEM – Northeast Peninsular Malay, SBR – Siberian Russia, THA – Thailand) .....	116

Figure 3.37 The tree of haplogroup Z. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, FIL – Philippines, JAP – Japan, SBR – Siberian Russia) .....	116
Figure 3.38 The tree of haplogroup M4'67. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (IND – India, PAK – Pakistan, SEM – Southeast Peninsular Malay, SWM – Southwest Peninsular Malay, THA – Thailand) .....	118
Figure 3.39 The tree of haplogroup M5. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (IND – India, NEP – Nepal, NWM – Northwest Peninsular Malay, PAK – Pakistan, XIN – Xinjiang) .....	119
Figure 3.40 The tree of haplogroup M24'41. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, NBO – North Borneo, VIE – Vietnam) .....	120
Figure 3.41 The tree of haplogroup M46 nested within M13'46'61. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (THA – Thailand, TMK – Thailand Moken).....	121
Figure 3.42 The tree of haplogroups M46 and M13'61. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. Certain branches of the tree are changed but some of the ML dates are kept on the tree. (CHI – China, IND – India, JAH – Semang Jahai, JAP – Japan, NEM – Northeast Peninsular Malay, NBO – North Borneo, NEP – Nepal, NWM – Northwest Peninsular Malay, SML – Aboriginal Malay Semelai, THA – Thailand, TIB – Tibet, VIE – Vietnam, XIN – Xinjiang, China) .....	122
Figure 3.43 The tree of haplogroup M2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BRA – Brazil, NWM – Northwest Peninsular Malay, PAK – Pakistan) .....	124
Figure 3.44 The tree of haplogroup M32. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AND – Andaman Islands, MAD – Madagascar, SEM – Southeast Peninsular Malay).....	125
Figure 3.45 The tree of haplogroup M47. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. ....	126
Figure 3.46 The tree of haplogroup M26. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, VIE – Vietnam) .....	127
Figure 3.47 The tree of haplogroup M20. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BID – Sarawak Bidayuh, CHI – China, JAV – Java, Indonesia, NBO – North Borneo, NEM – Northeast Peninsular Malay, Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, THA – Thailand, VIE – Vietnam) .....	129
Figure 3.48 The tree of haplogroup M51. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CAM – Cambodia, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NBO – North Borneo, NEM – Northeast Peninsular Malay, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUM – Sumatra, THA – Thailand, VIE – Vietnam) .....	130

Figure 3.49 The tree of haplogroup M50. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, VIE – Vietnam) .....	132
Figure 3.50 The tree of haplogroup M71. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, SEM – Southeast Peninsular Malay, THA – Thailand, VIE – Vietnam) .....	133
Figure 3.51 The tree of haplogroup M72. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, SWM – Southwest Peninsular Malay, VIE – Vietnam).....	134
Figure 3.52 The tree of haplogroup M42'74. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AUS – Australia, BID – Sarawak Bidayuh, CHI – China, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, SUM – Sumatra, VIE – Vietnam) .....	135
Figure 3.53 The tree of haplogroup M73'79. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, INA – Indonesia, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NBO – North Borneo, SUM – Sumatra, THA – Thailand, VIE – Vietnam).....	137
Figure 3.54 The tree of haplogroup M77. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (JAV – Java, Indonesia, NEM – Northeast Peninsular Malay, VIE – Vietnam).....	137
Figure 3.55 Schematic diagram of haplogroup D and its major subclades distribution. (AME – America, BRA – Brazil, CHI – China, EA – East Asia, FBT – Philippines Batak, FIL – Philippines, ISEA – Island Southeast Asia, JAP – Japan, NAS – North Asia, NWM – Northwest Peninsular Malaysia, THA – Thailand) .....	139
Figure 3.56 The tree of haplogroup D4a. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NWM – Northwest Peninsular Malay).....	140
Figure 3.57 The tree of haplogroups D5b and D5c. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan, SEM – Southeast Peninsular Malay, THA – Thailand).....	142
Figure 3.58 The tree of haplogroup D6. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, SBO – South Borneo) .....	143
Figure 3.59 The tree of haplogroups novel M* and M78 (Zhang et al., 2013). Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (SUL – Sulawesi, THA – Thailand, VIE – Vietnam) .....	143
Figure 4.1 Schematic diagram of haplogroup N's major subclades present in Southeast Asia. ....	145
Figure 4.2 Schematic diagram of haplogroup N9 and its major subclades. (EA – East Asia, SEA – Southeast Asia) .....	146

Figure 4.3 The tree of haplogroup N9a showing the subclades of N9a1'3, N2'4'5, N9a6, N9a7, N9a8, N9a9 and N9a10. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan) .....	146
Figure 4.4 The tree of haplogroup N9a6. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (JAH – Semang Jahai, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEL – Aboriginal Malay Seletar, TEM – Aboriginal Malay Temuan, SUM – Sumatra, BID – Sarawak Bidayuh) .....	147
Figure 4.5 Network of N9a* and N9a6 from HVS-I data. Figure adapted from Hill (2005). .....	148
Figure 4.6 The tree of haplogroup Y. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, FIL – Philippines, INA – Indonesia, JAP – Japan, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, TAI – Taiwan, XIN – Xinjiang) .....	150
Figure 4.7 HVS-I network of Y2. Figure adapted from Hill (2005). .....	150
Figure 4.8 The tree of haplogroup N21. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUM – Sumatra, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, THA – Thailand, VIE – Vietnam) .....	151
Figure 4.9 HVS-I network of N21. Denotation “Malay” is the Malay data used in Hill et al. (2006), “Malay ZZ” is the new data from Zafarina Zainuddin (personal communication). Figure adapted from Hill (2005). ...	152
Figure 4.10 The tree of haplogroup N22. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, LSI – Lesser Sunda Islands, SUM – Sumatra, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan) .....	153
Figure 4.11 HVS-I network of N22. Figure adapted from Hill (2005). .....	153
Figure 4.12 The tree of haplogroup N8. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAV – Java, SEM – Southeast Peninsular Malay, THA – Thailand, VIE – Vietnam) .....	155
Figure 4.13 The tree of haplogroup N10. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NEM – Northeast Peninsular Malay, SBO – South Borneo) .....	156
Figure 4.14 The tree of haplogroup N11. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. The Mamanwa sequences (Gunnarsdóttir et al., 2011a) with gaps are highlighted in red, and they were assumed not to carry private mutations in those gaps for the purpose of dating. (TIB – Tibet, CHI – China, MGL – Inner Mongolia, China, MAM – Philippines Mamanwa, and FIL – Philippines) .....	157
Figure 4.15 Schematic diagram of haplogroup A and its major subclades. (AME – America, EA – East Asia, NAS – North Asia) .....	158
Figure 4.16 The tree of haplogroup A5b and A5c. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, KOR – Korea, SEM – Southeast Peninsular Malay) .....	159

Figure 4.17 The tree of haplogroup A8 and A+152. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (SBR – Siberian Russia CHI – China, JAP – Japan, MGL – Inner Mongolia, China, RUS – Russia, IBE – Iberian Peninsula, AME – America, COL – Columbia).....	161
Figure 5.1 Schematic diagram of haplogroup R's major subclades present in Southeast Asia. ....	162
Figure 5.2 Schematic diagram of haplogroup B4 and its major subclades distribution. (EA – East Asia, NA – North Asia, NEU – North Eurasia, SEA – Southeast Asia and NO – Near Oceania) .....	163
Figure 5.3 The tree shows haplogroup B4a excluding B4a1a and B4a1c. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, VIE – Vietnam, BID – Sarawak Bidayuh, FIL – Philippines, KOR – Korea, TAI – Taiwan, SUM – Sumatra, SBO – South Borneo) .....	164
Figure 5.4 The tree of haplogroup B4a1c. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (JAP – Japan, SML – Aboriginal Malay Semelai, NWM – Northwest Malay, CHI – China, SBR – Siberian Russia MGL – Inner Mongolia, China, Tai - Taiwan).....	165
Figure 5.5 The tree of haplogroup B4a1a excluding B4a1a1. Time estimations shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAV – Java, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, MOL – Moluccas, NEM – Northeast Peninsular Malay, NBO – North Borneo, NWM – Northwest Peninsular Malay, SBO – South Borneo, SUL – Sulawesi, SUM – Sumatra, SWM – Southwest Peninsular Malay, TAI – Taiwan, THA – Thailand, PNG – Papua New Guinea) .....	165
Figure 5.6 The tree of haplogroup B4a1a1 excluding B4a1a1a1 and B4a1a1a4. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (BIS – Bismarck Island, BOU – Bougainville, COO – Cook Island, FIL – Philippines, JAV – Java, LSI – Lesser Sunda Islands, MIC – Micronesia, MOL – Moluccas, NWM – Northwest Peninsular Malay, SBO – South Borneo, SMO – Samoa, SUL – Sulawesi, TON – Tonga, PNG – Papua New Guinea, VAN – Vanuatu, WNG – West New Guinea).....	167
Figure 5.7 The tree of haplogroup B4a1a1a1 and B4a1a1a4. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (COO – Cook Island, PNG – Papua New Guinea, SMO – Samoa, VAN - Vanuatu) .....	167
Figure 5.8 HVS-I network of B4a1 (it is in fact B4a1a1a). Figure adapted from Hill (2005). ....	169
Figure 5.9 The tree of haplogroup B4g, B4h and B4i. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, TAI – Taiwan, THA – Thailand, VIE - Vietnam) .....	169
Figure 5.10 The tree of haplogroup B4b1 excluding B4b1a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, JAP – Japan, MGL – Inner Mongolia, China, RUS – Russia, VIE - Vietnam) .....	170
Figure 5.11 The tree of haplogroup B4b1a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. Sequences marked by “*” are erroneous and not used in age calculation. Sequences with “**” have np 310 removed as artefact since it forms incorrect evolutionary pathways in the clade. (CHI –	

China, FIL – Philippines, JAP – Japan, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, SBO – South Borneo, TEM – Aboriginal Malay Temuan) .....	170
Figure 5.12 HVS-I network of B4b1. Figure adapted from Hill (2005).....	172
Figure 5.13 The tree of haplogroup B4c1 excluding B4c1b2a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, JAP – Japan).....	173
Figure 5.14 The tree of haplogroup B4c1b2a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (FIL – Philippines, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, SWM – Southwest Peninsular Malay, TAI – Taiwan) .....	174
Figure 5.15 The tree of haplogroup B4c2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, KIN – Semang Kintak, SEL – Aboriginal Malay Seletar, SMI – Senoi Semai, SUM – Indonesia Sumatra, SWM – Southwest Peninsular Malay, THA – Thailand, UZB – Uzbekistan, VIE – Vietnam) .....	174
Figure 5.16 HVS-I network of B4c1b. Figure adapted from Hill (2005). .....	174
Figure 5.17 Schematic diagram of haplogroup B5 and its major subclades distribution. (EA – East Asia, MSEA – Mainland Southeast Asia, SC – Southern China, SEA – Southeast Asia).....	176
Figure 5.18 The tree of haplogroup B5a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BAT – Semang Batek, CAM – Cambodia, CHI – China, FIL – Philippines, JAP – Japan, NEM – Northeast Peninsular Malay, NIC – Nicobars, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, TAI – Taiwan, THA – Thailand, VIE – Vietnam) .....	178
Figure 5.19 The tree of haplogroup B5a2. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, TAI – Taiwan) .....	178
Figure 5.20 The tree of haplogroup B5b1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BAT – Semang Batek, CHI – China, FIL – Philippines, JAP – Japan, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, SEM – Southeast Peninsular Malay) .....	179
Figure 5.21 The tree of haplogroup B5b2 and B5b3. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BID – Sarawak Bidayuh, CHI – China, FIL – Philippines, JAP – Japan, SBR – Siberia, Russia).....	180
Figure 5.22 HVS-I network of B5a. Figure adapted from Hill (2005).....	180
Figure 5.23 HVS-I network of B5b. Figure adapted from Hill (2005).....	182
Figure 5.24 The tree of haplogroup R11. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NEM – Northeast Peninsular Malay, VIE – Vietnam) .....	182
Figure 5.25 The tree of haplogroup B6. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, NWM – Northwest Peninsular Malays, SEM – Southeast Peninsular Malay, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, VIE – Vietnam).....	183



Figure 5.26 The tree of haplogroup R12'21. DQ112752 is missing nps 16384-434 (Kivisild et al., 2006). Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AUS – Australia, BAT – Semang Batek, JAH – Semang Jahai, KEN – Semang Kensi, LAN – Semang Lanoh, NEM – Northeast Peninsular Malay, TMI – Senoi Temiar) .....	185
Figure 5.27 The tree of haplogroup R22. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, THA – Thailand, VIE – Vietnam).....	186
Figure 5.28 HVS-I network for R22 from HVS-I data. Figure adapted from Hill (2005). .....	186
Figure 5.29 Schematic diagram of haplogroup R9 and its major subclades distribution. (EA – East Asia, ISEA – Island SEA, JAP – Japan, MSEA – Mainland Southeast Asia, SAS – South Asia, SC – South China, SEA – Southeast Asia, TAI – Taiwan) .....	188
Figure 5.30 The tree of haplogroup R9b. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (ABM – Aboriginal Malay, JAH – Semang Jahai, JAV – Java, Indonesia, KIN – Semang Kintak, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, VIE – Vietnam).....	189
Figure 5.31 HVS-I network of R9b1 types. Figure adapted from Hill (2005). .....	189
Figure 5.32 The tree of haplogroup R9c1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, INA – Indonesia, FBT – Philippines Batak, FIL – Philippines, TAI – Taiwan).....	191
Figure 5.33 The tree of haplogroup F1a'c'f excluding F1a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (** – np 310 removed; BID – Bidayuh Sarawak, CHI – China, FBT – Philippines Batak, FIL – Philippines, JAP – Japan, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SUM – Sumatra, SWM – Southwest Peninsular Malay).....	192
Figure 5.34 The tree of haplogroup F1a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CAM – Cambodia, CHI – China, FIL – Philippines, JAH – Semang Jahai, JAK – Aboriginal Malay Jakun, JAP – Japan, NEM – Northeast Peninsular Malay, NIC – Nicobars, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, TAI – Taiwanese Aborigines, THA – Thailand, TMI – Senoi Temiar, TMK – Thailand Moken) .....	195
Figure 5.35 HVS-I network of F1a*. F1a is further defined by transitions at nps 16129 and 16172 (van Oven and Kayser, 2009). Figure adapted from Hill (2005). .....	195
Figure 5.36 HVS-I network of F1a1. Figure adapted from Hill (2005). .....	196
Figure 5.37 HVS-I network of F1a1a. Figure adapted from Hill (2005). .....	196
Figure 5.38 The tree of haplogroup F3. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FBT – Philippines Batak, INA – Indonesia, NBO – North Borneo, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, TAI – Taiwan, VIE – Vietnam, XIN – Xinjiang, China).....	198

Figure 5.39 The tree of haplogroup F4. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, IND – India, JAP – Japan, NWM – Northwest Peninsular Malay)	199
Figure 5.40 Schematic diagram of haplogroup P and its major subclades distribution. (AUS – Australia, FIL – Philippines, INA – Indonesia, PEM – Peninsular Malaysia, OCE – Oceania)	200
Figure 5.41 The tree of haplogroup P1+C16176T. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, NEM – Northeast Peninsular Malay, SWM – Southwest Peninsular Malay, PNG – Papua New Guinea)	201
Figure 5.42 HVS-I network for P1. Figure adapted from Hill (2005).	201
Figure 5.43 The tree of haplogroup R6. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (IND – India, NEM – Northeast Peninsular Malay, THA – Thailand)	202
Figure 5.44 The tree of haplogroup R7. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BRA – Brazil, IND – India, NBO – North Borneo, PAK – Pakistan, SWM – Southwest Peninsular Malay)	204
Figure 5.45 The tree of haplogroup R23 with three R* lineages. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (LSI – Lesser Sunda Islands, SWM – Southwest Peninsular Malay, THA – Thailand, VIE – Vietnam)	205
Figure 5.46 The tree of haplogroup U. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (ALG – Algeria, AMC – American, Caucasian, EUR – Europe, FRA – France, IND – India, ISR – Israel, ITA – Italia, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, PAK – Pakistan, RUS – Russia)	206
Figure 6.1 Bayesian skyline plot (BSP) indicating hypothetical effective population size over time of Orang Asli populations. The posterior effective population size through time is represented by the black line. The blue region represents the 95% confidence region. Effective population size is plotted on a log scale.	208
Figure 6.2 Bayesian skyline plot (BSP) indicating hypothetical effective population size over time of Peninsular Malay populations. The posterior effective population size through time is represented by the black line. The blue region represents the 95% confidence region. Effective population size is plotted on a log scale.	209

## List of Tables

<i>Table 2.1 Distribution of three Orang Asli subgroups from Peninsular Malaysia including samples from Hill et al., (2006), K.C. Ang and my Orang Asli samples. ....</i>	<i>50</i>
<i>Table 2.2 22 pairs of nested primers used for complete mtDNA genome PCR amplification designed and optimised by Maria Pala (research fellow in the Archaeogenetics Research Group, Huddersfield).....</i>	<i>54</i>
<i>Table 2.3 23 pairs of alternative nested primers used for complete mtDNA genome PCR amplification. ....</i>	<i>55</i>
<i>Table 2.4 32 pairs of alternative nested primers used for complete mtDNA genome PCR amplification (Maca-Meyer et al., 2001).....</i>	<i>56</i>
<i>Table 2.5 Primers for sequencing reactions, designed and optimised by Maria Pala. (Tm – annealing temperature).....</i>	<i>58</i>
<i>Table 2.6 The general settings for BEAUTi v1.7.4. ....</i>	<i>68</i>
<i>Table 3.1 mtDNA haplogroup frequencies of 85 Orang Asli from Semang and Senoi populations. ....</i>	<i>71</i>
<i>Table 3.2 mtDNA haplogroup distribution of 18 Orang Asli subgroups (K.C. Ang). ....</i>	<i>76</i>
<i>Table 3.3 Combined Orang Asli subgroups and haplogroups for mtDNA HVS-I of 260 samples in Hill et al. (2006), 91 K.C. Ang (personal communication) and 85 (this study). ....</i>	<i>78</i>
<i>Table 3.4 Distribution of the modern Malay samples grouped according to sample regions and haplogroups. The four regions in Peninsular Malaysia are Northeast Peninsular Malay (NEM), Northwest Peninsular Malay (NWM), Southeast Peninsular Malay (SEM), and Southwest Peninsular Malay (SWM).....</i>	<i>79</i>
<i>Table 7.1 Assignment of Orang Asli lineages to putative proximal source regions. Figures taken from Table 3.3. (EA – East Asia, ISEA – Island Southeast Asia, MSEA/SUN – Mainland Southeast Asia/Sunda) ....</i>	<i>212</i>
<i>Table 7.2 Assignment of Malay lineages to putative proximal source regions. Figures taken from Table 3.4. (EA – East Asia, ISEA/NG – Island Southeast Asia/New Guinea, MSEA/SUN – Mainland Southeast Asia/Sunda, SAS – South Asia).....</i>	<i>223</i>

## List of Abbreviations

AD	Anno Domini
AMHs	Anatomically modern humans
ATP	adenosine triphosphate
BC	Before Christ
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BI	Bayesian Inference
bp	base pairs
BP	Before Present
BSP	Bayesian Skyline Plot
COI	cytochrome c oxidase subunit I
COII	cytochrome c oxidase subunit II
COIII	cytochrome c oxidase subunit III
CSB	conserved sequence block
D-loop	displacement loop
DNA	Deoxyribonucleic acid
dNTPs	deoxynucleoside triphosphates
EDTA	ethylenediaminetetraacetic acid
G6PD	glucose-6-phosphate dehydrogenase
HPD	Highest posterior density
HSP	heavy-strand promoter
HVS-I	hypervariable segment I
HVS-II	hypervariable segment II
HVS-III	hypervariable segment III
ISEA	Island Southeast Asia
ka	thousand years
LGM	Last Glacial Maximum
LSP	light-strand promoter
MCMC	Markov Chain Monte Carlo
ML	Maximum likelihood
MP	Maximum Parsimony
MRCA	most recent common ancestor
mRNA	messenger RNA
MSEA	Mainland Southeast Asia
MSY	Male-specific region of the Y chromosome
mtDNA	mitochondrial DNA
NJ	Neighbour-Joining
np	nucleotide position
OA	<i>Orang Asli</i>
O <sub>L</sub>	Light-strand origin of replication
OXPHOS	Oxidative phosphorylation

PAML	Phylogenetic Analysis by Maximum Likelihood
PCR	Polymerase chain reaction
RBC	red blood cells
rCRS	the revised Cambridge Reference Sequence
RD	Reduced-median
RFLP	restriction fragment-length polymorphism
RNA	Ribonucleic acid
ROS	Reactive oxygen species
rpm	revolutions per minute
rRNA	ribosomal RNA
RSRS	reconstructed sapiens reference sequence
SDS	Sodium dodecyl sulphate
SE	Standard error
SEA	Southeast Asia
SNPs	single nucleotide polymorphisms
STRs	Short tandem repeats
TBE	tris-borate-EDTA
tRNA	transfer RNA
ya	years ago

# 1 Introduction

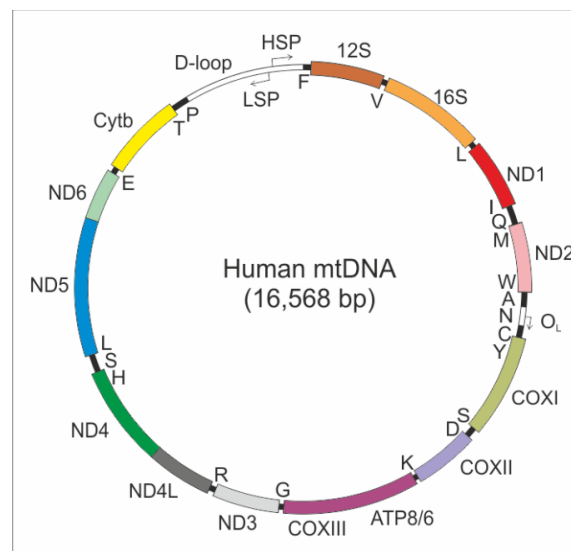
## 1.1 Mitochondria

Mitochondria are one of the cytoplasmic organelles present in eukaryotic organisms. Their main function is generating energy in the form of ATP (adenosine triphosphate) through the process of oxidative phosphorylation. Other functions include intracellular signalling and involvement in apoptosis, intermediate metabolism, antiviral responses and the metabolism of amino acids, lipids and nucleotides (Cruz *et al.*, 2005; Chinnery, 2006; McBride *et al.*, 2006). Margulis first proposed in the 1960s (see Margulis, 1981), and it is now accepted, that mitochondria originated from endosymbiotic bacteria, which were taken up into eukaryotic cells about 1.5 billion years ago.

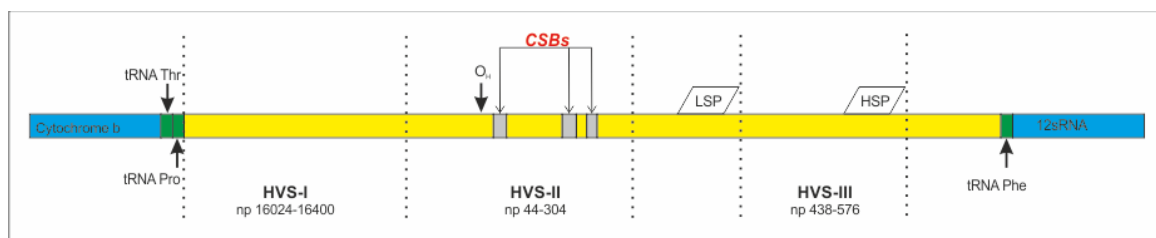
Mitochondria contribute a non-recombining region of genetic material, a circular double-stranded DNA molecule 16,568 base pairs in length in humans (Anderson *et al.*, 1981; Andrews *et al.*, 1999). The mammalian mtDNA contains 13 protein coding genes of the respiratory chain, two ribosomal RNAs (rRNA) and 22 transfer RNA (tRNA) genes (Figure 1.1). The two strands of mtDNA are named heavy strand (H) and light strand (L) respectively depend on their density in guanine content. There are very few introns between the protein-coding genes and the genes occasionally overlap with each other's. There is a small non-coding region of around 1,100 bp called control region or D-loop, which is crucial for replication and transcription. The control region consists of three hypervariable segments (HVS-I, II and III) (Figure 1.2).

Each mammalian cell contains mtDNA ranging from 1,000 to 10,000 copies depend on the energy requirement. These copies could all be identical copies of mtDNA (homoplasmy) or some may have a mutation in them while the rest do not (heteroplasmy). Heteroplasmy could present as somatic mutations or be passed on to the offspring if it is in the germline. As a haploid genome, mtDNA is maternally inherited as a single molecule unchanged to the next generation, unless new mutations occur. It is now commonly accepted that mtDNA does not recombine during meiosis. Previously, different reports claimed to have found evidence for recombination in mitochondrial DNA. However, they were quickly dismissed for different

reasons, such as sequencing error (Hagelberg *et al.*, 1999); error was introduced in the transcription of the data and confusion caused by sites scoring (Macaulay *et al.*, 1999b); the data used in the linkage disequilibrium study for recombination were erroneous (Kivisild *et al.*, 2000); the use of inadequate statistical tools (Awadalla *et al.*, 1999; Herrnstadt *et al.*, 2002a; 2002b; Ingman *et al.*, 2000) as reported by Piganeau and Eyre-Walker (2004); or simply, results obtained by chance (Innan and Nordborg, 2002).



**Figure 1.1 Schematic diagram of the human mitochondrial genome. The genome encodes 22 transfer RNAs (indicated by single letter abbreviation) between the coding genes, two ribosomal RNAs (12S and 16S), and 13 essential genes that encode subunits of the oxidative phosphorylation enzyme complexes. The D-loop region includes heavy and light-strand promoters (HSP and LSP), and the origin of L-strand replication ( $O_L$ ).**



**Figure 1.2 Diagram of the mammalian mtDNA control region (Modified from Chinnery, 2006). The control region consists of three hypervariable segments (I, II and III) separated by three conserved sequence blocks (CSBs). LSP and HSP are the promoters for the L and H- strand respectively.**

Sperm mitochondria sometimes manage to enter the oocyte during fertilization, but they are quickly degraded before implantation by ubiquitin-dependent process that eliminates the sperm in the oocyte. There is only one case of paternal inheritance of mtDNA reported in a patient with a myopathy condition (Schwartz and Vissing, 2002), however it is so rare that it can be ignored.

Fortunately, a major advantage of using network analysis in phylogenetic study of mtDNA is that it is able to highlight potential recombination between lineages in the form of reticulations and pinpoint sample mix-up (Bandelt *et al.*, 2004). Bandelt *et al.* (2001) used network analysis to identify possible errors found in phylogenetic analysis and classified them into five classes which helped to lower the error rate and obtain more precise results (see 2.7.2). When it was seemingly detected, the recombination was most likely due to sequencing errors that generated the systematic artefacts of the phantom mutations (Bandelt *et al.*, 2002). Furthermore, so far there are no such results reported in the literature of complete mitochondrial sequences available online.

The observation of non-maternal inheritance of mtDNA was reported in the muscle tissues of a Danish patient suffering from a mitochondrial myopathy (Schwartz and Vissing, 2002). However this singular case so far was neither confirmed in another lab nor could be found in other cases of sporadic myopathies (Filosto *et al.*, 2003; Taylor *et al.*, 2003; Schwartz and Vissing, 2004). Bandelt *et al.* (2005) re-analysed the data including other cases and reported that the phenomenon of mixed or mosaic mtDNA can be concluded as contamination and sample mix-up.

Homoplasmic mtDNA mutations are transmitted to all maternal offspring, but the transmission of heteroplasmic mtDNA is more complex. Homoplasmy is when all copies of the mitochondrial genome are identical; heteroplasmy is when there is a mixture of two or more mitochondrial genotypes (Taylor and Turnbull, 2005). There is a rapid intergenerational change in mitochondrial genotypes and levels of heteroplasmy during maternal mitochondrial transmission, which appears to be governed by random genetic drift. This change, however, does not present in the segregation of mutant mtDNA in both non-dividing and proliferating tissues. The nature of the genetic drift (a reduction in genetic diversity resulting from a population bottleneck) during oogenesis is not completely known, and the amount of mutated mtDNA that is transmitted to the offspring is variable (Brown *et al.*, 2001; Taylor and Turnbull, 2005). It could be due to a physical reduction in the number of mitochondrial genomes within individual cells, a reduction in the effective population size due to the compartmentalisation of genomes into homoplasmic segregating units, or the preferential amplification of specific genotypes (Chinnery, 2006; Cree *et al.*, 2008; Khrapko, 2008).



### **1.1.1 Mutation rates**

Early studies of human and other primate DNAs revealed that the base substitution rate in mtDNA is about ten times faster than the average rate in the nuclear genome (Brown *et al.*, 1979). The faster rate could be due to several reasons; firstly, the mtDNA has a less robust repair mechanism when mutational events occur and is unable to correct it efficiently. Secondly, the mtDNA molecule is less protected by proteins than the nuclear genome, and lastly, it is in close contact with reactive oxygen species (ROS) generated during oxidative phosphorylation (OXPHOS), which are the main cause of mutations in mtDNA (Fernandez-Silva *et al.*, 2003).

### **1.1.2 Studies using human mtDNA**

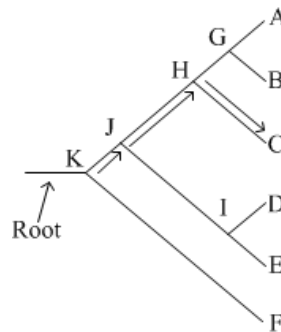
Archaeogenetics is the application of molecular genetics to the study of the human past. It is especially concerned with the reconstruction of the dispersal history of humankind. The study mostly relies on DNA data from living populations, as well as a substantial contribution from ancient DNA studies. Phylogeography is the study of the geographical distribution of the lineages in a phylogeny (Avise *et al.*, 1987; Jobling *et al.*, 2003). The underlying principle is that every mutation takes place at a particular point in space and time, and each event can be in theory reconstructed from the distribution. In other words, phylogenetic analysis is applied to geographically labelled samples where it is possible to estimate the number and timings of different colonisation events using the geographic origin of the samples and the time depth of lineages on the genealogical tree. However, this approach is not always straightforward due to the low density of mutations and recombination in autosomes and the X chromosome, and high drift in the Y chromosome. Since mutations occur frequently in mtDNA, the high density of mutations tracks the distribution of lineages through space and over time in higher detail and precision especially when complete mtDNAs are used. By comparing the mtDNA lineages from one region to another, it is possible to infer the direction and timing of dispersals.

There are four main components in phylogeographic analysis:-

#### **1.1.2.1 The gene tree or network**

A gene tree or network represents the hierarchical history between the hypothetical basal lineage and its descendants. It is formed by branches and nodes. External (terminal) nodes represent the extant taxa, often called operational taxonomic units (OTUs). Internal

nodes are called the hypothetical taxonomic units (HTUs) to indicate that they represent the ancestral taxa. A group of taxa (Figure 1.3, the taxa A, B, and C) that share the same branch with similar set of haplotypes (mutational differences) have a monophyletic origin and is called a clade, or haplogroup. C, D and E do not form a clade since a clade includes all descendants of a common ancestor (that would therefore include A and B), they are called paraphyletic. The branching order of the nodes is called the topology of the tree (Vandamme, 2009).



**Figure 1.3 Structure of a rooted tree. A, B, C, D, E, and F are external nodes. G, H, I, J, and K are internal nodes, with K as root node. In a rooted tree, the arrow indicates the direction of evolution. Figure adapted from Vandamme (2009).**

Both mtDNA and the Y chromosome are markers commonly used in phylogeographic studies because they are haploid, and non-recombining. Without recombination, the differences between lineages will be solely derived from mutations and the order of the mutation will trace the history of the locus (Macaulay and Richards, 2001).

#### **1.1.2.2 The geographic distribution of lineages**

The geographic distribution of lineages is assessed by sampling at different locations and then identifying similar lineages between locations. Subsequently, to reflect the extent of the geographic distribution of the lineages, the published mtDNA genomes of related lineages (GenBank, FamilyTree or Genome Projects) from adjacent locations are incorporated into the phylogeny.

#### **1.1.2.3 The application of a molecular clock**

A molecular clock is applied in the phylogeographic analysis to infer the time depth of the lineage of the phylogenetic tree. Earlier molecular clock analyses assumed that the diversity accumulated at a linear rate (Bromham and Penny, 2003; Kumar, 2005), which was problematic. Recently, the molecular clock was corrected for purifying selection for the entire

mtDNA molecule and calibrated with recent evidence for the divergence time of humans and chimpanzees (Soares *et al.*, 2009). See more details in 1.2.2.

#### **1.1.2.4 Other complementary lines of evidence**

The phylogeographic analysis works best with a model-based framework that uses complementary evidence from other fields including archaeology, linguistics, climatology, geology, palaeontology and radiocarbon dating. Genetic data cannot alone serve as a predictor for the cultural and linguistic affiliation of its carrier. In other words, phylogenetic analysis is able to show the magnitude of immigration at a particular point of time and location, but there is nothing in the genetic evidence *per se* that will associate the two (Richards *et al.*, 2002).

Radiocarbon dating by  $^{14}\text{C}$  started in the 1950s and continues to be the most widely employed method of inferring chronometric age for late Pleistocene and Holocene age materials (Taylor, 1995). The two main methods employed in radiocarbon dating are decay counting methods (using liquid scintillation of acid-washed or acid-base-acid (ABA), and gas proportional counters) and accelerator mass spectrometry (AMS). AMS requires small sample size and it may be possible to use a pre-treatment method (such as acid-base-oxidation (ABOX) for charcoal, or ultra-filtration for bone) that cannot be applied while retaining a large sample size (Ramsey, 2008). Alternative dating techniques include thermoluminescence and optical dating. Optical dating was used to estimate the time since the quartz sediments were last exposed to sunlight (Huntley *et al.*, 1985; Aitken, 1998).

## **1.2 The Human mtDNA Phylogenetic Tree**

Early human mtDNA studies were performed using RFLP analysis (restriction fragment-length polymorphism) of the whole genome and control-region sequencing. The first human complete mtDNA sequence was published by Anderson *et al.* (1981) and later the revised Cambridge reference sequence (rCRS) by Andrews *et al.* (1999). Since then, rCRS became the reference sequence for scoring the polymorphisms present in mtDNA sequences and for building the human mtDNA phylogenetic tree. Recently, Behar *et al.* (2012) proposed a replacement sequence to rCRS with a newly reconstructed basal sequence called the Reconstructed Sapiens Reference Sequence (RSRS). It was inferred by rooting the mtDNA phylogenetic tree with the Neanderthal complete mtDNA sequence. However, the rCRS is so

well established that it is both troublesome and potentially highly confusing to change to the new RSRS system with a risk of introducing errors; here I used the widely accepted rCRS system.

### **1.2.1 mtDNA tree nomenclature**

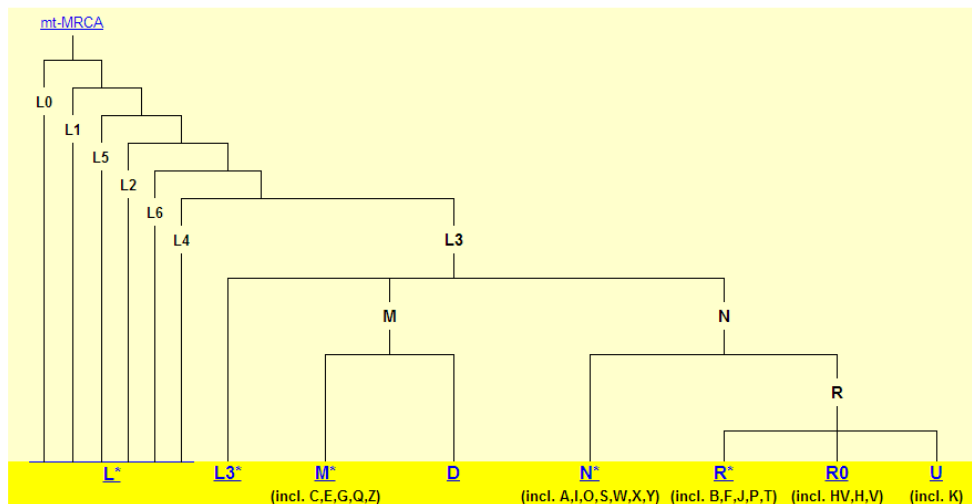
The present nomenclature of mitochondrial clades was introduced by Torroni *et al.* (1993) with the four main Native American haplogroups, A, B, C and D. Subsequently, new haplogroups were discovered and soon took up all the alphabet letters. Even though it was found in high diversity in Africa, the African lineages were assigned to one single letter L because they were collectively and erroneously placed as one single haplogroup with an Asian lineage as outgroup (Chen *et al.*, 1995). In fact, L should subsume the whole modern human mtDNA tree.

The nomenclature follows a simple order that represents the nesting structure of clades and subclades within the tree. Consider a hypothetical haplogroup V, diverged into two branches; they will be labelled as V1 and V2. Two subclades of V1 would be V1a and V1b, and two subclades of V1a are V1a1 and V1a2, and so on. Despite the issue with haplogroups L0, L1 etc., re-labelling the whole mtDNA phylogenetic tree would be both impractical and confusing to the scientific community. To avoid changing of existing labels when a more ancestral node is discovered, new ancestral node can be labelled with a prefix “Pre-”. The star symbol “\*” is used to represent members of a haplogroups that do not yet belong to a defined subclade, essentially defining paraphyletic clusters. Using an example from above, any other less frequent and undefined V lineages apart from V1 and V2 will be labelled as V\*.

A tree of global human mtDNA based on complete genome variation including haplogroup nomenclature was published by van Oven and Kayser (2009) on [www.phylotree.org](http://www.phylotree.org). The online tree was last updated to mtDNA Tree Build 15 version (30 Sep 2012) with a total of 16,810 complete human mtDNA sequences (Figure 1.4).

### PhyloTree.org - mtDNA tree Build 15 (30 Sep 2012)

For your convenience, the online tree is divided into eight subtrees accessible through the links in the scheme below. Alternatively, the entire tree can be downloaded as single file: [mtDNA tree Build 15 \(30 Sep 2012\)](#)



PhyloTree collaborates with [EMPOP](#) and [mtDNACommunity](#) to enrich our knowledge of the human mtDNA phylogeny in its evolution and present-day variation.

Maintained by Mannis van Oven ([m.vanoven@erasmusmc.nl](mailto:m.vanoven@erasmusmc.nl))  
Department of Forensic Molecular Biology  
Erasmus MC, University Medical Center Rotterdam  
The Netherlands

**Figure 1.4** The simplified view of the global human mtDNA phylogenetic tree Build 15 version (30 Sep 2012). The root of the tree (mt-MRCA) represents the maternal most recent common ancestor of all humans. Haplogroups L0 to L6 are lineages specific to Africa, except that L3 diverged into M and N from which all the remaining diversity is derived. For more details, see: [www.phylotree.org](http://www.phylotree.org), van Oven and Kayser (2009).

The mitochondrial tree was traditionally rooted using chimpanzee, bonobo, and gorilla sequences (Hixson and Brown, 1986; Arnason *et al.*, 1996; Xu and Arnason, 1996). Cann *et al.* (1987) analysed 147 human mtDNA drawn from five geographic populations using RFLPs, they found the lineages coalesced to a common ancestor who was postulated to have lived about 200,000 years ago in Africa. She was popularly known as “Mitochondrial Eve”, the ancestral woman from whom all of modern humans were descended or the female line of descent. The age of mitochondrial Eve was nearly 200,000 ya estimated by a mutation rate derived from the coding region only (Mishmar *et al.*, 2003), 186,000 ya with synonymous rate (synonymous transition rate of 1 per 7884 years by Soares *et al.* (2008)); the synonymous mutation per 6764 years reported by Kivisild *et al.* (2006) was too fast), and about 190,000 ya with whole mtDNA genome corrected for purifying selection (Soares *et al.*, 2008). In 2008, Green and collaborators published the Neanderthal complete mitochondrial genome (Green *et al.*, 2008), and the first split was found to occur between haplogroup L0 and all of the remaining haplogroups both within and outside Africa.

### 1.2.2 The mtDNA molecular clock

Several mutation rates have been proposed for the human mtDNA genome. Forster *et al.* (1996) estimated a value one transitions per nucleotide per 20,180 years for the HVS-I control region, and Mishmar *et al.* (2003) suggested one transition per 5138 years based on the coding region np 577-16023, which is more than 10 times lower. These mutation rates assumed a clock-like evolution for the human mtDNA with a homogenous distribution of the mutation rate across time. However, the mtDNA phylogeny shows higher proportions of non-synonymous coding mutations at the tips of the branches than deeper in the tree, indicating purifying selection is acting progressively on mtDNA (Kivisild *et al.*, 2006; Pereira *et al.*, 2011). Kivisild *et al.* (2006) proposed, using a phylogeny consisting of 277 individuals, a mutation rate of one transition in 6884 years for synonymous substitutions only. Soares *et al.* (2009) provided a revised estimate of the synonymous substitutions rate, and found that it is considerably slower at one synonymous mutation per 7884 years. Synonymous substitutions are not under selection pressure and present lower saturation compared with the control region. All these estimates utilized a divergence time of 6 or 6.5 million years between chimpanzee and *Homo sapiens*.

Endicott and Ho (2008) attempted a recalibration with Bayesian estimation but assuming several highly debatable intraspecific calibration points, inevitably introducing assumptions into the estimation. Their Bayesian approach assumed a relaxed molecular clock for the coding and non-coding regions of the mtDNA, but they did not provide a rate for the whole mtDNA molecule. Henn *et al.* (2009) made the first attempt to characterise the mutation-rate curve of human mtDNA-coding region, but this suffers from the same weakness regarding calibration-point assumptions and did not allow for selection.

However, the observed mutation rate is non-uniform throughout the mtDNA molecule. This is most likely due to purifying selection on the mtDNA where different regions are under different selective constraints. Even within the fast-evolving control region, the mutation rate is not homogeneous, especially the nucleotide positions identified as hotspots by Endicott *et al.* (2009). Overcoming the concerns from previous reports, Soares *et al.* (2009) constructed a phylogenetic tree with more than 2000 complete genomes and assumed a single calibration point corresponding to the chimpanzee-*Homo sapiens* split. They generated the mutation rate for the complete mtDNA genome, and estimated  $1.665 \times 10^{-8}$  substitutions per nucleotide per year, or one mutation every 3624 years. Various classes of

mutations at different phylogenetic time depths were estimated and used to correct the mutation rate in each temporal window. Separate clocks for the synonymous mutations and the non-coding segments were also estimated for comparison with the previous studies.

### 1.3 Y chromosome haplogroup phylogeny

The Y chromosome contains the largest non-recombining block in the human genome and is therefore very useful for evolutionary/population studies, forensics, medical genetics, and genealogical reconstruction. Similar to mtDNA phylogenies, the MSY (male-specific region of the Y chromosome) phylogeny has nomenclature for designated haplogroups, and the branches are defined by SNPs (single-nucleotide polymorphisms) (de Knijff, 2000).

The root of the MSY tree has been coined as the Y-chromosome Adam, who is paternally the most recent common ancestor (TMRCA) with an estimated time ~200 ka (Francalacci *et al.*, 2013). Mendez *et al.* (2013) claims a much older time ~338 ka using an African American sample that was found to add an ancient root to the MSY tree. According to Mendez *et al.* (2013), the older age was estimated with a slower mutation rate ( $4.39 \times 10^{-9} - 7.07 \times 10^{-9}$  per base per year) obtained from the whole-genome sequence data (Roach *et al.*, 2010; Conrad *et al.*, 2011; Kong *et al.*, 2012), although they also estimated TMRCA to ~209 ka when they utilized the higher mutation rate ( $1.0 \times 10^{-9}$  per base per year) (Cruciani *et al.*, 2011). The deepest primary splits in the Y chromosome phylogeny are African-specific clade A, and clade BT, the latter gives rise to clades B and CT. Both A and B are restricted to Africa, and CT comprises the majority of African and all non-African chromosomes (Underhill and Kivisild, 2007a; Karafet *et al.*, 2008; Cruciani *et al.*, 2011). Haplogroup BCT diverged around 75 ka into two subclades, haplogroups B and CT (including DE), the latter migrated outside of Africa until recent times. Haplogroup DE is distributed in Africa (E) and Asia (D). Haplogroup C is widely found in South and East Asia, Oceania, and North America. In East Asia, the most frequent lineage is haplogroup K which further diverged into haplogroups N and O (Underhill and Kivisild, 2007a). The topology of the MSY phylogenetic tree, along with the geographical distribution of the major clades A, B, and CT, has been interpreted as supporting an African origin for AMH (Underhill *et al.*, 2000), with the deepest lineages found in Khoisan of south Africa and Ethiopians of east Africa (Hammer *et al.*, 2001; Semino *et al.*, 2002).

Sex-specific dispersals happened when different mating and migration routes exist within the populations, e.g. the Jewish people (Behar *et al.*, 2004; Behar *et al.*, 2010). Besides, different genes are subject to different selective forces. Therefore, it is not unusual for the two uniparentally inherited marker systems to occasionally provide evidence for different evolutionary histories within the same geographic regions (Carvajal-Carmona *et al.*, 2000; Oota *et al.*, 2001; Destro-Bisol *et al.*, 2004; Bolnick *et al.*, 2006). However, there are broad consistent features between the phylogenies derived from both mtDNA and Y-chromosome (Underhill and Kivisild, 2007a).

- Both phylogenies support the African root because only African populations harbour lineages of both the primary branches descended from the root of the phylogenies;
- A small subset of both mtDNA and Y chromosome trees is distributed outside Africa. The non-African founder lineages are haplogroups M, N and R of mtDNA, and C, D and F of Y chromosome (Kivisild *et al.*, 2003);
- Australia and Europe show limited founder composition compared to Asia
- Some recent gene flow such as the Bantu (Cruciani *et al.*, 2002; Salas *et al.*, 2002; Luis *et al.*, 2004) and Polynesian expansions (Hage and Marck, 2003; Kayser *et al.*, 2008; Soares *et al.*, 2011), have left traces in the genetic composition of both markers;
- Admixture was observed in regions such as North Africa, and Central Asia, as well as West Asia, where intermediate variation is seen between distinctive pools of mtDNA and Y chromosome varieties (Wells *et al.*, 2001; Arredi *et al.*, 2004; Comas *et al.*, 2004; Quintana-Murci *et al.*, 2004).

Earlier analyses by Capelli *et al.* (2001) studied the paternal heritage of the Austronesian-speaking peoples of SEA and Oceania. They found that the majority of current Austronesian speakers trace their paternal heritage to Pleistocene settlers in the region, contrary to models arguing for replacement more-recent agricultural immigrants (Bellwood, 1997). A fraction of the paternal heritage, however, traced to more-recent immigrants from northern 'Neolithic' populations. They also found some northern Neolithic component dispersed throughout the region, with a higher contribution in SEA and a nearly complete absence in Melanesia. Later on, Karafet *et al.* (2010) found the paternal gene pool is sharply divided between western and eastern locations, along the Wallace's line between the islands



of Bali and Flores. Karafet and colleagues found that the eastern Y chromosome haplogroups are closely related to Melanesian lineages which were likely to reflect the initial colonisation of the region, while the majority of western Y chromosome haplogroups may have entered Indonesia during the Palaeolithic from MSEA.

Black *et al.* (2006) studied eight Y chromosome binary polymorphisms of the Cambodian population by comparing them with other populations from Bali, Christmas Island, Malaysia, Miao people, Southern Han and Northeastern Thai, and they found there is a dominant East Asian male ancestry throughout SEA. The Cambodian community has been reported to display post-Neolithic East Asian male and pre-Neolithic Southeast Asian female ancestries similar to those reported in other Southeast Asian populations (Su *et al.*, 2000; Karafet *et al.*, 2005; Wen *et al.*, 2005).

Cai *et al.* (2011) studied the Mon-Khmer and Hmong-Mien speaking populations in South China and MSEA, and found that a predominant MSY haplogroup O3a3b-M7, dating ~19 ka, showed an early unidirectional distribution from SEA into East Asia, which was suggested to result from the genetic drift of East Asian ancestors carrying O3a3b-M7 lineages through many small bottlenecks complicated by landscape between SEA and East Asia. He *et al.* (2012) analysed the Y-chromosome variations of the Cham people and showed that while there are indigenous components in both MSY and mtDNA markers, there are also indications showing genetic admixture, presumably from Austronesian-speaking-immigrants from ISEA with the local populations in MSEA, as well as some Y chromosome influences from South Asia.

Simonson *et al.* (2011) compared the MSY of Austronesian-speaking Iban population in Sarawak, Malaysia with individuals from East and Southeast Asia populations. The MSY haplogroup frequencies show male-specific gene flow from SEA, and the admixture analysis and PCA illustrate a similar pattern of population differentiation, with the Iban population showing affinity to those from MSEA and Indonesian samples. However, they were not able to preclude more recent but less substantial contributions from northern populations such as those of Taiwan. Delfin *et al.* (2011) studied the Filipino populations including the negrito groups, where they found heterogeneity present in both negrito and non-negrito groups with signatures of old and recent periods, and diverse affinities. They identified two Y-chromosome haplogroups C-RPS4Y and K-M9 predominant among the Filipino negritos which represent founding lineages in the Asia-Pacific region that are also shared with

indigenous Australians, and not found among the Filipino non-negrito populations. Hence, they conclude a possible divergence and subsequent gene flow between some Filipino negrito groups and indigenous Australians, not necessarily via direct contact between these groups, but gene flow from Australia to the Philippines via neighbouring populations, in a ‘stepping-stone’ manner, although they admitted that additional loci are needed to confirm the signal (Delfin *et al.*, 2011).

## 1.4 Autosomal Markers

The genome-wide autosomal DNA variations or markers serve as another line of evidence to help understand the population genetic ancestry in relation to linguistic, geographic and demographic history. The newer genome-wide autosomal approach assay a huge number of autosomal single-nucleotide polymorphisms (SNPs) using genechips. A large-scale study by the HUGO Pan-Asian SNP Consortium (Abdulla *et al.*, 2009) on East Asian (EA) and SEA populations showed that more than 90% of EA autosomal variation could be found in either SEA or Central-South Asian populations and show clinal structure with haplotype diversity decreasing from south to north, indicating that SEA was a major geographic source of EA populations. Abdulla *et al.* (2009) also tested the two-wave hypothesis that ancestral negrito populations first settled in SEA, Australia, and Oceania before a northerly migration originating in or near the Middle East, and spreading both towards Europe and Northeast Asia via Central Asia (Cavalli-Sforza *et al.*, 2003). Their results do not disprove the two-wave model; instead, they found a population history that unites the negrito and non-negrito populations of SEA and East Asia via a single primary wave of entry of AMHs into the continent (Abdulla *et al.*, 2009). The study also included several Filipino negrito groups, where they found no clear-cut genetic distinction between the Filipino negrito and non-negrito groups. This conclusion seems at odds with the MSY (haplogroups C-RPS4Y and K-M9) and mtDNA data (haplogroups B4b1a2) indicating novel and ancient haplogroups in the Filipino negrito groups, as mentioned by Delfin *et al.* (2011). A possible explanation given by Delfin *et al.* (2011) is that the ancestors of the Filipino negrito groups were isolated from the ancestors of the non-negrito groups, but then the two groups have experienced substantial, primarily male-mediated admixture in recent times (Stoneking and Delfin, 2010). Such substantial admixture, recorded in the MSY data, could

possibly account for the partial results reflected by the genome-wide SNP data (Abdulla *et al.*, 2009).

Other autosomal analyses in SEA showed a certain amount of admixture (~20%) between presumed-Austronesian-speaking and non-Austronesian-speakers prior to further eastward migration of the presumed-Austronesian migrants (Friedlaender *et al.*, 2008; Kayser *et al.*, 2008). The autosomal SNPs study by Jinam *et al.* (2012) claims that the Malaysian negrito, Philippine negrito and Alorese in Indonesia are distributed individually apart in a gradient or comet-like pattern in their PCA result, suggesting recent admixture between these groups with Thai, Chinese, or other Austronesians who formed a tight cluster (see Figure 4 in Jinam *et al.*, 2012).

Hatin *et al.* (2011) detected genetic substructure of four Malay sub-ethnic groups of Peninsular Malaysia (*Melayu Kelantan*, *Melayu Minang*, *Melayu Jawa* and *Melayu Bugis*) using genome-wide SNPs. The *Melayu Minang*, *Melayu Jawa* and *Melayu Bugis* are all known recent (i.e. historic) colonies of settlers from known parts of Indonesia. The studies indicate the existence of genetic heterogeneity in these populations that relate to their diverse origins and recent histories. The neighbour-joining tree showed, as expected, that the *Melayu Jawa*, *Melayu Bugis* and *Melayu Minang* formed a cluster with Indonesian populations indicating a common ancestry, while the *Melayu Kelantan* formed a distinct group indicating they are genetically different from the other Malay sub-ethnic groups (Hatin *et al.*, 2011), which is consistent with an unpublished heuristic assignment of origins, based on the HVS-I data in Nur Haslindawaty *et al.* (2010) compared with our Southeast Asian database (this assignment was performed by S. Oppenheimer, personal communication).

Wong *et al.* (2013) contributed to population genetics studies in SEA by whole-genome sequencing at a minimum of 30x coverage of Malay samples from Singapore to characterise the polymorphic variants in the population. The studies reported that they have detected deep population-level rare and low-frequency variants among the Austronesian-speaking Malay, which were not found in other populations by the low-pass sequencing either in the 1000 Genomes Project (1KGP; McVean *et al.*, 2012) or the International HapMap Project (2003) that did not include the Malay.

Both mtDNA and MSY markers have been particularly subject to the effects of random genetic drift, and each autosomal marker, no matter how informative, still represents a minute fraction of the total genetic variation among populations (Kayser *et al.*, 2008). The detailed

analyses of genealogical lineages and in particular the sex-dependent demographic scenarios that lie behind them can only be ascertained and dated, at present, through uniparental marker systems. Although, more autosomal data from additional populations combined with demographic modelling are required to sort out the relative roles of residence pattern, society structure, amount of admixture, and subsequent migration and drift in shaping the autosomal, mtDNA and MSY gene pools (Friedlaender *et al.*, 2008; Kayser *et al.*, 2008).

## **1.5 The Origin of Anatomically Modern Humans (AMH)**

Many of the potential archaeological sites are at present most likely submerged under the sea since the beachcombing course taken by the modern humans is dependent on a seashore environment. Between 70 ka and today, an ~80 m-rise of sea level has altered the coastline by shifting it hundreds of kilometres inland and potentially inundating the range of beachcombing AMHs (Metspalu *et al.*, 2006). Other factors that also prevent the recovery of archaeological and palaeontological evidence include the tectonic movements of the continental shelves and the accuracy of fossil dating techniques which is in constant dispute (Chen and Zhang, 1991; Klein, 1999).

Most hominin fossils older than 100 ka are outside the scope of molecular genetics because they do not contain enough DNA to analyse, except for the Neanderthals where the more recent fossils still contain endogenous DNA (Goodwin and Ovchinnikov, 2006). One group of archaic hominins, the Neanderthals, evolved in Europe for around 200,000 years before they became extinct about 30 ka. They were then suggested to be replaced by anatomically modern humans (AMHs) called the Cro-Magnons, which started in the east around 45 ka in Europe and later retreated to several refugia 30-28 ka (Goodwin and Ovchinnikov, 2006). Two main hypotheses have been proposed to explain the emergence of AMH – the multiregional evolution hypothesis and the Out of Africa hypothesis.

The multiregional hypothesis argues that AMH evolved from the archaic populations (*Homo erectus* ranging from Africa to Asia and Europe) in different continents they had spread to, simultaneously with innovations to help cope with new environments. According to this hypothesis, the AMH was the result of continuous, parallel development from *Homo erectus* to *Homo sapiens*, homogenizing the differences between them, preventing speciation events (Wolpoff *et al.*, 1988). Multiregionalists claim that regional features can be followed through the fossil record to modern humans with no need for an influx of Africans, and the

morphological similarities shared between modern human and Neanderthals were the direct evidence that modern humans received them directly from Neanderthals as this appears more than likely the trait evolving twice (Wolpoff and Thorne, 1991). For this to happen would depend on the occurrence of numerous migrations and interbreeding between populations from different parts of the world. There are no obvious parallels to this in other species of other animals in different continents (Ayala, 1995). The main alternative to continuous multiregional evolution would be the Out of Africa replacement hypothesis that does not require the concept of parallel evolution.

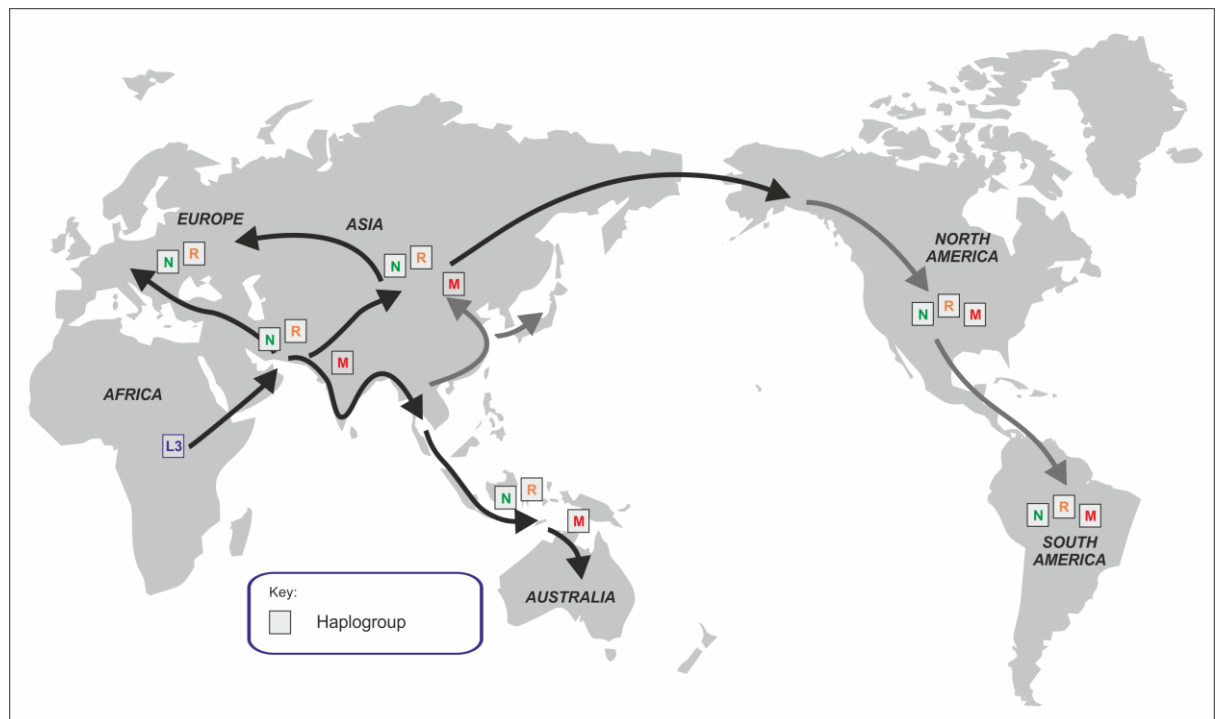
## 1.6 Out of Africa

Palaeontological evidence shows that the late Pleistocene fossils from China resemble European and African middle Pleistocene hominins more than their supposed local ancestors. The earliest *Homo sapiens* fossils are also found in Africa and near the Levant; no clear Neanderthal/*Homo sapiens* transitional fossils have been found in Europe despite the excellent fossil record, and modern humans seem to have been present in the Levant briefly before Neanderthals (Stringer and Andrews, 1988). Ponce de Leon and Zollikofer (2001) argued that Neanderthals and modern humans were separate species based on cranial examination. *Homo floresiensis*, discovered recently in Flores, Indonesia, lived until as recent as 18,000 ya. A miniaturised species with very small brains, they most probably evolved from *Homo erectus* without gene exchange with other hominins, and may have no contact with *Homo sapiens* (Brown *et al.*, 2004; Morwood *et al.*, 2004).

The oldest modern human remains were found in southern Ethiopia. The human cranial remains were dated by feldspar crystals found below the fossil levels with  $^{40}\text{Ar}/^{39}\text{Ar}$  method to around 198 ka (McDougall *et al.*, 2005). In Herto, Ethiopia, the AMH fossils were dated by the associated fossils and artefacts with  $^{40}\text{Ar}/^{39}\text{Ar}$  method to 160-154 ka (Clark *et al.*, 2003; White *et al.*, 2003). Modern human remains dated between 80 to 100 ka were also found to occupy sites in South Africa: with dental remains from Blombos cave (Grine and Henshilwood, 2002) and fragmentary artefacts from Klasies River Mouth. It is therefore difficult to pinpoint the origin of AMH within Africa (Deacon and Geleijnse, 1988; Deacon, 1992).

Earlier mtDNA work appeared to support the out of Africa model, but these early studies suffered problems due to their lack of resolution. A reanalysis of the original mtDNA

data (Cann *et al.*, 1987) by Templeton (1993) claimed there were other more parsimonious trees than the original. Some of the more parsimonious trees carried an African-only branch as did the original, while others had a mixed African-Asian primary branch. This did not prove that the MRCA was non-African; it showed that while more parsimonious trees existed, a more complete approach of analysis was needed. These problems have since been largely overcome by the use of more extensive sampling, combining control-region sequencing with coding-region RFLP typing (Torroni *et al.*, 1996; Macaulay *et al.*, 1999a), complete mtDNA sequencing (Ingman *et al.*, 2000; Maca-Meyer *et al.*, 2001; Herrnstadt *et al.*, 2002a; Behar *et al.*, 2008) and the use of better phylogenetic analytical tools (Bandelt *et al.*, 1995; Penny *et al.*, 1995; Yang, 1997; Drummond and Rambaut, 2007).



**Figure 1.5** Map showing the inferred movements of mtDNA haplogroups in Africa and Eurasia between 60 to 30 ka. The figure indicates the African exodus on the Southern route ~ 60 ka, taking the coastal path along the Indian Ocean into Southeast Asia and Australasia. Figure modified from Family Tree DNA (2006), Soares *et al.* (2009) and Beyin (2011).

Ancient mtDNA has also lent its support to the out of Africa hypothesis. The analyses of the mtDNA control-region sequence of the Neanderthal specimens found in western Germany (Krings *et al.*, 1999) and Mezmaiskaya Cave in the northern Caucasus (Ovchinnikov *et al.*, 2000) appeared to fall outside the variation of modern humans and suggested that the Neanderthal mtDNAs and the AMHs mtDNA gene pool have evolved as separate entities for a substantial period of time. The estimated date of divergence between

the mtDNAs of the Neanderthal and modern humans by control-region has been estimated, for instance, as ~465 ka (Krings *et al.*, 1999). In recent years, multiple complete ancient mtDNAs were able to be sequenced with the advancement of new sequencing technologies. Comparison of Neanderthal and modern human mtDNAs has indicated that the divergence time between the two can be estimated at around 511-550 ka (Briggs *et al.*, 2009; Soares *et al.*, 2009).

The most recent common coalescent ancestor, also called “Mitochondrial Eve”, lies at the root of the two most basal branches in the tree, L0 and the remaining human mtDNA lineages L1 through L6 (Figure 1.4). MtDNA haplogroups L0 to L6 are found in sub-Saharan African populations and constitute the deepest branches of the global human mtDNA tree, indicating an African origin for *Homo sapiens* mtDNAs at about 200,000 years ago. All non-African mtDNA lineages form subclusters of the African clade L3 that expanded from East Africa approximately 60 ka (Mountain *et al.*, 1995; Watson *et al.*, 1997). L3 left Africa and diverged into haplogroups M and N (Forster, 2004). Recently, Soares *et al.* (2009) published a calibrated molecular clock using the complete mtDNA genome with a maximum likelihood approach and estimated the age of clade L3 at ~70 ka. Non-African sub-lineages of L3: M and N radiated towards Asia, a small subset of N lineages colonised Eurasia and Europe. After the Last Glacial Maximum (LGM) ~19 ka, the first wave of a set of founders (A2, B2, C1, D1 and X2a) entered America from the north and spread across the American continent from north to south, following the Pacific coastal route (Howell *et al.*, 2003; Achilli *et al.*, 2008).

## 1.7 The Southern Route

The Southern Route model is currently the main model for the earliest modern human colonisation of Asia and is supported by many recent genetic, archaeological, and anthropological studies (Lahr and Foley, 1994; Quintana-Murci *et al.*, 1999; Stringer, 2000; Kivisild *et al.*, 2003; Oppenheimer, 2003; Kivisild *et al.*, 2004; Metspalu *et al.*, 2006). When comparing data from different geographic locations (South Asia, East Asia, Australia), it is evident that each region carried different sub-branches of M, N and R descending directly from the root of the three haplogroups (Ingman and Gyllensten, 2003; Kong *et al.*, 2003b; Palanichamy *et al.*, 2004; Friedlaender *et al.*, 2005; Merriwether *et al.*, 2005; Thangaraj *et al.*, 2005; Kong *et al.*, 2006; Sun *et al.*, 2006; Thangaraj *et al.*, 2006; van Holst Pellekaan *et al.*,

2006). Macaulay *et al.* (2005) found the presence of basal M, N and R lineages in the *Orang Asli* groups in Peninsular Malaysia suggesting these three founders moved along the south coast of Asia ~50-60 ka, reaching Southeast Asia and the Sahul continent (Australia and New Guinea) by ~50 ka (Stringer, 2000; Mellars, 2006; Underhill and Kivisild, 2007; Shi *et al.*, 2010; Fernandes *et al.*, 2012; Soares *et al.*, 2012). However, different continental regions each harbour a distinct set of basal descendants of the M, N and R. This suggested that each region was colonised by individuals primarily carrying the root type of the three founders in their mitochondrial gene pool and that differentiation occurred in each region as mutations accumulated independently locally in these new M, N and R subclades (Macaulay *et al.*, 2005; Metspalu *et al.*, 2006). Haplogroups M and N effectively cover the whole mtDNA pool in Asia. Haplogroup M is slightly more frequent than N in Siberia, northern China, Japan, and South Asia, while in Southeast Asia it is the opposite. M is nearly absent from Southwest Asia, where they are mainly subsets of N (mostly R) in the mtDNA pool. The colonisation of Europe would have been the results of an early offshoot of the southern route out of Africa, involving only lineages from the two founders N and R (Macaulay *et al.*, 2005). The estimated ages of M and N in East Asia, Southeast Asia, and Australia, and the slightly higher age of N and R in South Asia, support a single rapid dispersal out of Africa that took place within the last 70 ka (Macaulay *et al.*, 2005; Metspalu *et al.*, 2006).

Earlier works focused on migration routes into China and East Asia, and also out into Polynesia. Ballinger *et al.* (1992) suggested that South China was the centre of modern human expansion in East Asia based on the higher levels of mtDNA diversity found in the south. This was subsequently supported by work done on the Y chromosome that found the Northern Asian ancestry can be traced to the south, which additionally strengthens the case for a northward migration of modern humans into eastern Asia after ~60 ka (Su *et al.*, 1999).

An alternative model of earlier modern human colonisation of southern Asia has been proposed recently. According to the model, modern humans dispersed from Africa at a much earlier time before 74 ka reaching southern Asia before the catastrophic volcanic eruption of Mount Toba in Sumatra at ~74 ka (Oppenheimer, 2003; Clarkson *et al.*, 2009; Haslam *et al.*, 2010). The event in Mount Toba was the largest eruption in the past 2 million year, producing dense rock equivalent volume of 800 km<sup>3</sup> of ash into the atmosphere that blanketed the skies and blocked out sunlight for six years (Ambrose, 1998). As a result, global temperatures dropped to colder than during the Last Glacial Maximum 19-25 ka, with some suggesting that



this caused a middle Pleistocene human population bottleneck at this time. However, the pre-Toba artefacts dated to more than 125 ka from the Jurreru Valley in South Asia are no longer thought to be the handiwork of modern humans but most likely the work of archaic people (Appenzeller, 2012). Even with the existing mtDNA database of 16,810 complete sequences, there are no pre-L3 non-African mtDNA found to indicate an early modern human colonisation of South Asia before the Toba eruption (Appenzeller, 2012; Mellars *et al.*, 2013).

## **1.8 Southeast Asia**

Southeast Asia (SEA) consists of Mainland Southeast Asia (MSEA) and Island Southeast Asia (ISEA). MSEA includes the present day Myanmar, Thailand, Cambodia, Vietnam, Laos and Peninsular Malaysia. ISEA includes East Malaysia, Brunei, Indonesia (excluding West Papua), East Timor and the Philippines. SEA has a tropical climate, the day temperature floats around 30 °C throughout the year and it lies in a zone of high humidity with large areas under the regime of the monsoonal system (Verstappen, 1997). It covers an area from latitude 20 ° north and 16 ° south, and longitude 95 ° west to 105 ° east. As the climate on the equator does not change much, the biodiversity of flora and fauna is high because of the effect of climate and geological history (Myers *et al.*, 2000). Soils in SEA tend to be infertile clays, where most nutrients are cycled within the rainforest biomass rather than in the topsoil. There are few edible wild plants suitable for human consumption or animals that are dispersed or arboreal and difficult to hunt in equatorial forests (Bellwood, 1994).

### **1.8.1 The flooding of Sundaland**

The palaeoenvironmental and palaeogeographic changes during the late Pleistocene were crucial in shaping the population history of Sundaland. Intermittent periods of global warming and cooling during the last ice ages resulted in glaciation and deglaciation of the Arctic ice caps. The alternating rise and lowering of the sea levels had a profound impact on the climate and biogeography of the region. Nowhere else in the world has experienced such a large-scale loss of landmass as a consequence of rise in sea levels. The land area lost by Sundaland after the Ice Age was as large as India (Oppenheimer, 1998).

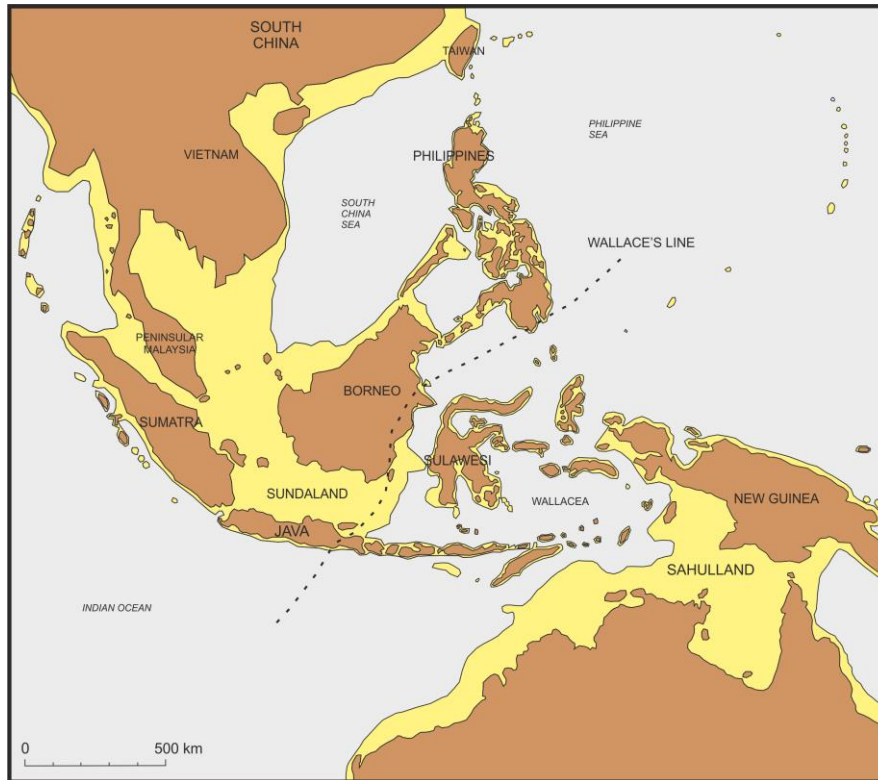
The Last Glacial Maximum (LGM) occurred around 19-25 ka, centered ~21 ka, and is marked by the maximum volume of ice sheet expanded over Scandinavia and northern

Europe (Banks *et al.*, 2008). Glacio-eustatic depression of sea level by ~120 m at the LGM exposed the Sunda shelf joining MSEA to Sumatra, Java, Borneo and possibly Palawan, substantially reducing the size of the South China Sea (Bird *et al.*, 2005). This exposed continent shown in Figure 1.6 was called “Sundaland” (Molengraaff, 1921), which was essentially a south-east extension of the continental shelf of Southeast Asia; similarly Taiwan was a peninsular of the Chinese coast. East of Sundaland was a boundary called Wallace’s Line separating Asia by water from Wallacea and the Sahul Shelf, despite the low sea levels. Alfred Russell Wallace was among the first naturalists to observe a clear distinction between eastern and western faunas across Wallace’s Line, and recognised a natural barrier for the spread of mammals, including early hominins (Wallace, 1881; Voris, 2000). The sea-level rises that began ~19 ka due to early warming in the oceans led to the drowning of the shallow landmasses, losing almost half of the land area, thus revealing the present day geographical appearance. Current geographical features of Southeast Asia were formed towards the end of the Pleistocene epoch, where the Sunda flooding became stabilised ~6-7 ka.

After the LGM, the ice retreat started between 19.5-16 ka caused a climatic improvement and vast areas were exposed to be re-settled reaching a peak between 16-14 ka (Terberger and Street, 2002; Gamble *et al.*, 2004). After the rapid warming, the climate experienced a fast cold snap to glacial conditions called the Younger Dryas ~11.5 ka, which are most probably due to the cold melt-waters that invaded the Atlantic Ocean caused by global warming and the melting of the ice sheets. After the Younger Dryas, the climate warms up reaching the optimal conditions for a widespread growth of wild cereals and legumes (Scarre, 2005).

In Southeast Asia, the ancestral population spread on east along the coast, with successive founder effects amplifying the founder lineages of haplogroups M, N and R as they moved, diverging in mainland Southeast Asia and the prehistoric continent of Sundaland (now the Indo-Malaysian archipelago) and continued onwards into eastern Eurasia and Australasia from the coast (Richards *et al.*, 2006). Based on the ages of haplogroups of M, N and R, and the absent of human settlements evidence before ~30 ka, Richards *et al.* (2006) suggested that the crossing into Australia and Papuans would probably follow the easier ‘northern route’ from Sundaland, via Wallacea (rather than the ‘southern route’ via the Nusa Tenggara). Nevertheless, they also noted that the water crossing would have been shorter by either way at any time during this period than it is today. Alternatively, a dispersal model

suggested by Bird *et al.* (2005) claimed that around 110-85 ka and thereafter around 70 ka, land bridges emerged intermittently connecting the Asian mainland with Sumatra, Java and Borneo. An open vegetation savannah corridor crossing the interior of Sundaland probably became an inland route for the first modern human dispersals throughout much of the region and into Australia (Bird *et al.*, 2005).



**Figure 1.6 Sundaland in the Late Pleistocene period. Areas in yellow were drowned when the sea-level rose; brown areas indicate the present day countries. Figure adapted from Oppenheimer (1998).**

During the last glacial period, about half of Sundaland was flooded when a meltwater pulse originating in the northern hemisphere, probably caused by early warming in the oceans of the southern hemisphere and leading to a rise in sea-levels. The immersed lands are now 70-80 m below present day sea levels (Pelejero *et al.*, 1999; Clark and Mix, 2002). Two main episodes of sea-level rise were identified between 15-13.5 ka and 11.5-10 ka (Blanchon and Shaw, 1995; Pelejero *et al.*, 1999), and on top of that, Blanchon and Shaw (1995) detected another episode between 8-7 ka. These floodings after the LGM were related to an increase in the sea-surface temperature and an increase in marine productivity (Pelejero *et al.*, 1999). Earthquakes and tsunamis were likely to take place when the earth's crust needed to rapidly readjust to the new distribution of water and ice in the sea (Oppenheimer, 1998).

Subsequently, it is claimed that towards the end of the Pleistocene and early Holocene, a number of large terrestrial animals became extinct throughout the planet (Louys *et al.*, 2007). Particularly in ISEA, the increasing sea-level also affected the structure of vegetation, disrupting river systems, and human populations in the region (Voris, 2000).

### **1.8.2 First settlement of ISEA by modern humans**

The fossil and archaeological evidence of hominin occupation in SEA showed that *Homo erectus* was present in Java as early as 1.6 million years ago (Swisher *et al.*, 1994) and may have lived until as late as 27 ka (Swisher *et al.*, 1996). These dates were arguably unreliable because of the uncertainties with the dating method and the stratigraphic position of the Java fossils (Roberts *et al.*, 2005). There is no archaeological evidence for the arrival of modern humans to Southeast Asia prior to ~50 ka. Several widely accepted archaeological dates of modern human occupation are obtained from the “Deep Skull” in Niah Cave of Sarawak, Borneo, radiocarbon dated to around 39-45 ka (Barker *et al.*, 2005), the Jerimalai shelter in Wallacea ~42 ka (O’Connor *et al.*, 2007), New Guinea 44-49 ka (Summerhayes *et al.*, 2010), the Bismarck Archipelago ~33 ka (Allen *et al.*, 1988), and the Northern Solomon Islands ~28 ka (Wickler and Spriggs, 1988).

In Niah Cave, the evidence of biomass burning suggested that humans occupied the location since at least 50 ka (Hunt *et al.*, 2007). In MSEA, evidence of Pleistocene modern humans was discovered in Lang Rongrien, southwestern Thailand, radiocarbon dated between 27 to 43 ka (Anderson, 1990, 1997). Up north in China, in Tianyuan Cave of Zhoukoudian, the oldest modern human remains were dated to 39-42 ka (Shang *et al.*, 2007). The dates seem to imply that ISEA was *en route* to the ancient continent of Sahul (Australia and New Guinea) and the time of colonisation of ISEA should predate the time of colonisation of Sahul. <sup>14</sup>C dating method suffers the technical limits of chronometry to around 50 ka, hence becoming meaningless effectively when this is exceeded. However, the time of colonisation in Australia has so far been controversial, with some of the highest age estimates for modern humans outside of Africa, an age of 60 ka obtained from thermoluminescence method (Fullagar *et al.*, 1996). However, O’Connell and Allen (2004) argued that estimated ages older than 45 ka obtained from the thermoluminescence method were not well-supported, due to the method’s lack of sensitivity to archaeological context.

The time of earlier colonisation by modern humans in Sundaland has been difficult to estimate because a number of potentially archaeological sites may well be now submerged as a result of sea-level fluctuations. The earlier inhabitants of Sundaland are likely to have resided along the coast and exploited marine resources. Bellwood (1997) suggested that for over a timespan of at least 40 ka, ISEA was the ultimate source-region for the populations of Australia and the Pacific Islands. Thangaraj *et al.* (2005) studied the mtDNA variation in the relict aboriginal populations from the Andaman Islands (Indian Ocean), and obtained an estimated age for the M lineages of ~65 ka. In the same year, Macaulay *et al.* (2005) estimated an age of ~63 ka for both haplogroups M and N in the *Orang Asli* from Peninsular Malaysia, although the estimated ages were slightly overestimated (Soares *et al.*, 2009). Mellars *et al.* (2013) reported that their genetic evidence from both Africa and Asia and the archaeological evidence from South Asian sites have found it unlikely that the initial dispersal of AMHs from Africa to southern Asia occurred before the volcanic eruption of the Mount Toba volcano at ~74 ka. These studies, however, indicate a single and rapid southern coastal route from Africa, along coastal India to Southeast Asia, although the estimated ages were slightly overestimated (Soares *et al.*, 2009; Mellars *et al.*, 2013).

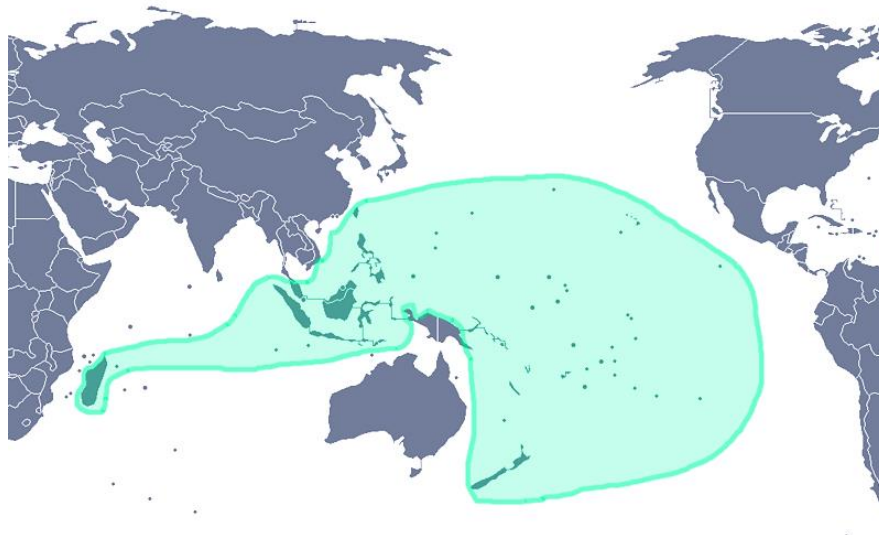
### **1.8.3 The People of SEA**

The population history of the indigenous people of Southeast Asia has been vigorously debated since the 1970s. Various theories were hypothesised on the basis of archaeological finds and linguistic surveys in attempts to explain the origins and patterns of prehistoric human dispersal in the region. In Southeast Asia, the first settlements of the so-called “Australo-Melanesian” or negrito populations are traditionally considered to have arrived from the Horn of Africa during the Pleistocene period via a “southern route” around 60 ka to 75 ka, along coastal India to Southeast Asia ~50 ka and Australasia (Lahr and Foley, 1994; Cavalli-Sforza *et al.*, 1994; Bellwood, 1997; Turney *et al.*, 2001; Soares *et al.*, 2009) before the sea-level rises. The negrito populations are present in the Semang of Peninsular Malaysia, resembling the Andaman Islanders and Filipino Aeta in that they are short in stature with dark skin and woolly hair. Bellwood (1997) also proposed a second group that migrated later from the northern latitudes called the “Mongoloids”. They represent almost all of the rest of the populations in ISEA who speak the Austronesian languages.

The Pacific Islands consists of Melanesia, Micronesia and Polynesia. They are not part of SEA but they are archaeologically and linguistically closely related to ISEA. The term

Australo-Melanesian broadly includes the Melanesians (from the New Guinea highlands and Island Melanesia), Australians and the negrito populations in SEA (Bellwood, 1997). The people of Melanesia and the New Guinea highlands are dark-skinned, and the Melanesian populations speak Papuan languages and are considered to be descendants of the first settlers in the region. On the other hand, both the Micronesians and Polynesians have lighter skin and shared similar cultural and linguistic background, speaking the Austronesian languages (Terrell, 1986; Terrell *et al.*, 2001).

The Austronesian languages are spoken throughout ISEA, except for some populations in Eastern Indonesia who speak Papuan languages. The Austronesian-speaking groups have a common ancestral language, Proto-Austronesian, with approximately 1200 Austronesian languages estimated. They are spread across ISEA, distributed as far west as Madagascar, to the northern coast of New Guinea and the Pacific Islands in the east (Pawley, 2002). The Papuan-speaking (non-Austronesian) groups lack a recent common ancestry and include numerous linguistically unrelated groups (Wurm and Hattori, 1981; Specht, 2005). These two groups of people have a different history in Melanesia. The non-Austronesian speakers reflect the early Pleistocene arrival of modern humans in the region, whereas Austronesian speakers were thought to have arrived as migrants from East Asia by 3.5 ka (Kirch, 1997). The extent of the Austronesian languages family is shown in Figure 1.7.



**Figure 1.7** The area highlighted shows the extent of the Austronesian migrations. Figure adapted from Quirino (2010).

Palaeoanthropological analysis includes osteoscopy examination of non-metric traits such as sexual dimorphism and population affinity characteristics from the cranial and

postcranial bones. The osteometry examination assesses the metric traits from the skull and teeth, from a number of populations and analysed by statistical tools to ascertain the relationship between the various populations. Studies on osteometry in Asia have suggested that the populations from ISEA cluster closely with those from MSEA. The Polynesian population forms a separate branch between Southeast Asia and Melanesia, and does not appear to be affiliated with the Taiwan and China populations which cluster together (Pietrusewsky, 1997; Matsumura and Hudson, 2004; Hanihara and Ishida, 2005).

Turner (1987) studied the SEA populations using dental morphological traits, and suggested that two migrations originated from central China ~20-30 ka, which can be represented by two set of dental features, the Sinodonts and Sundadonts. The Sundadonts are generally found in the south that exhibits a pattern of simplification and retention. The Sundadonts showed weaker expression for traits like incisor shovelling, double-shovelling, four-cusps lower molars, and retained ancestral traits such as two-rooted upper first premolars and two-rooted lower second molars. The Sinodonts are found in the north that shows intensification and addition. They have more pronounced grades of shovelling, double-shovelling, three-rooted lower first molars and peg-shaped upper third molars. The Sinodonts expanded northward into China, Siberia and across the Bering land bridge into America. The Sundadonts moved southward into Southeast Asia and Indonesia, and later through Melanesia, Micronesia and Polynesia (Turner, 1987).

Admixture analysis with autosomal SNPs which are highly informative for Asian-Melanesian ancestry carried out by Cox *et al.* (2010) showed that the East Indonesians display a clinal transition from Asian to Melanesian genetic variants along the Wallace's Line. This phenotypic gradient probably reflects mixing of two long-separated ancestral source populations – one descended from the initial Melanesians, and the other related to the arrival of Palaeolithic immigrants in ISEA and/or with the spread of agriculture. They also noticed a high signal of Asian X-linked markers throughout the transition zone, which seems to suggest that the admixture process was sex-biased, either signalling a westward expansion of patrilocal Melanesian groups or an eastward expansion of matrilocal Asian inhabitants. The observed sex bias in admixture rate may be due to the matrilocal residence system that dominated the ancestral Austronesian societies (Cox *et al.*, 2010).

Soares *et al.* (2008) carried out complete mtDNA genome sequencing of haplogroup E, a lineage with important mtDNA diversity in the region. They showed that it has evolved *in*

*situ* over the last 35 ka. It then expanded around the beginning of the Holocene throughout ISEA, which coincides with the post-Last Glacial Maximum (LGM) sea-level rises that broke up the Sundaland continent into present day archipelago. There were at least three major bursts of accelerated sea-level rise and flooding in the so-called Catastrophic Rise Events 1 to 3, possibly due to ice sheet collapse, at ~14.5, 11.5, and 7.5 ka (Blanchon and Shaw, 1995). Therefore, it is suggested that most probably the postglacial climate change and sea-level rises around 15 – 7 ka were the main forces shaping modern human dispersals in the region instead of farming/language model. Soares *et al.* (2008) also mentioned that haplogroup E lineages are associated with the “flake-blade technocomplex”, an industry based on flakes detached from rotated multiplatform cores, which emerged around 25 – 30 ka restricted to the islands and coastlines of the Sulu Sea region. Around 18 ka, this distinctive stone tool technology spread to northern Borneo and throughout ISEA by the maritime-oriented populations living around the coastlines at that time.

#### **1.8.4 The “Out of Taiwan model” and the “Farming/Language Dispersal Hypothesis”**

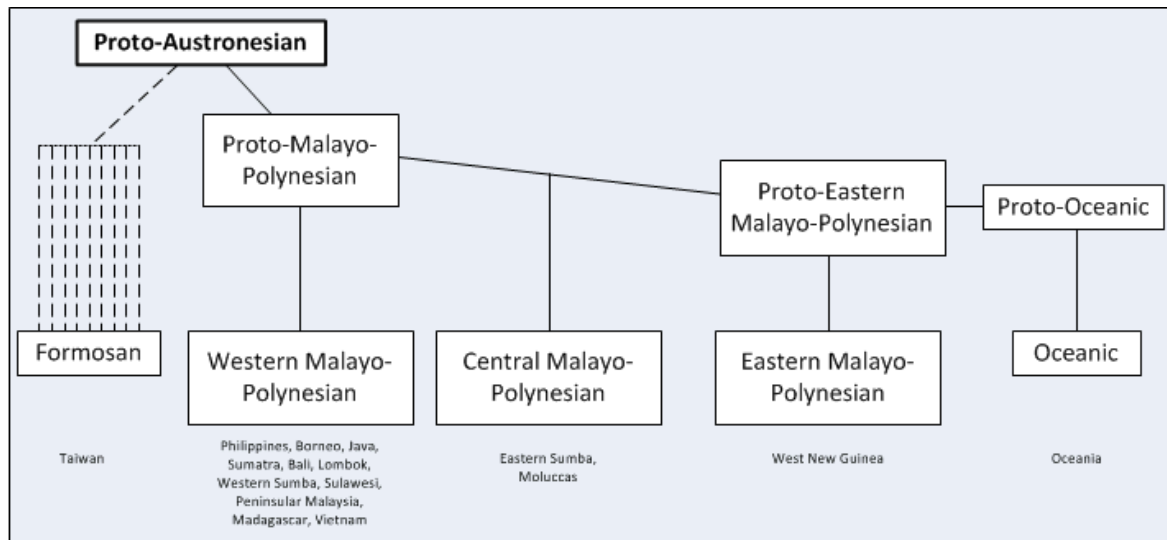
One dominant model has explained the colonisation process of ISEA, which has different names over the years: the “Express train to Polynesia” (Diamond, 1988), “Out-of-Taiwan” (Bellwood, 1997) and the “Farming/Language Dispersal hypothesis” (Bellwood and Renfrew, 2003). The out-of-Taiwan or express train to Polynesia model recognises two waves of migration by two groups of people, the initial “Australo-Melanesian” or “Australoid” and the “Mongoloid” who arrived later. These first settlements were then replaced or assimilated in the mid-Holocene by a maritime driven dispersal of Austronesian speakers from southern China via Taiwan into ISEA (Bellwood, 1997, 2005a, 2005b; Diamond and Bellwood, 2003; Bellwood and Dizon, 2008). According to this model, rice-agriculturalists speaking proto-Austronesian languages migrated from southern China ~5,500 ya reaching Taiwan, before expanding into the Philippines and the rest of ISEA ~4,000 ya. Diamond and Bellwood (2003) identified three main advantages of agricultural populations over the hunter-gatherer populations: (i) higher food production per area leading to possible higher population density; (ii) a sedentary lifestyle that can accumulate stored food surpluses, which were a prerequisite for the development of complex technology, social stratification, centralized states, and professional armies; and (iii), the farming populations acquired



resistance against epidemics originating from domestic animals, for e.g. smallpox and measles.

Fundamentally the out-of-Taiwan model is one of the groups of models that incorporate the broader “Farming/Language Dispersal Hypothesis” (Bellwood and Renfrew, 2003). The 15 language families are: (i) Bantu (Niger-Congo family), (ii) Arawak (Taino), (iii) Austro-Asiatic, Tai or “Daic” and Sino-Tibetan, (iv) Uto-Aztecan, (v) Oto-Manguean, Mixe-Zoquean, Mayan, (vi) New Guinea Highlands, (vii) Japanese, (viii) Austronesian, (ix) Dravidian, (x) Afro-Asiatic, and (xi) Indo-European. These language families have all been related to nine different homelands: “hearths” of agriculture, or centres of domestication (Diamond and Bellwood, 2003). Bellwood (2001) outlined several characteristics that indicate the spread of a linguistic family together with a farming expansion, (i) a set of lexical reconstructions related to crops and domestic activities; (ii) rapid spread of a language over a large area; (iii) linguistic time depths corresponding to the archaeological records of the Neolithic event; and (iv) elements that allow particular languages to be linked with particular archaeological material cultures and a sudden horizon-like appearance of the cultures archaeologically.

There are five main phyla of languages spoken in East Asia: (i) Sino-Tibetan in China, Burma and Nepal; (ii) Hmong-Mien in South China, North Vietnam and Laos; (iii) Tai-Kadai in South China and Indochina; (iv) Austro-Asiatic in Indochina and Central Malaysia; and (v) Austronesian in ISEA, Taiwan and much of the Pacific (Sagart *et al.*, 2005). Blust (1995) suggested the Austronesian language family based on the occurrence of shared innovations in phonology and pronoun forms. There are possibly nine main branches of Austronesian languages in Taiwan, which are collectively called the Formosan by Blust (1995) (Figure 1.8). The decomposition and structuring of the Malayo-Polynesian branch can be used to map the expansion of the language in ISEA. The Malayo-Polynesian branch can be divided into the following subgroups: Western Malayo-Polynesian (spoken in the Philippines, Borneo, Sulawesi, Java, Bali, Lombok, West Sumbawa, Sumatra, Peninsular Malaysia, Vietnam, and Madagascar), Central-Malayo-Polynesian (in the Eastern Sumba and Moluccas except Halmahera), and Eastern-Malayo-Polynesian (spoken in South Halmahera, West New Guinea, Melanesia, Micronesia and Polynesia) (Blust, 1995).



**Figure 1.8 Austronesian languages with corresponding geographical location. Figure after Blust (1995).**

The model suggests that the Austronesian languages originated after the colonisation of Taiwan by Neolithic pottery-making and rice-growing farmers in southern China and Taiwan after 4,000 BC, in a cultural environment of increasing population density, advancing technology (including boat construction and carpentry), and increasing dependence on agriculture and animal domestication, and also a portable food production repertoire that allowed long-distance dispersal to take place (Bellwood, 2011). This was followed by the spread of farming, pottery, and Neolithic tools to replace or hybridize with the original population of the Philippines (2,000 to 1,500 BC), then southwest to the Malay Peninsular and to Madagascar, and east through Indonesia out across the Pacific to the furthest islands of Polynesia, eventually reaching New Zealand by about 1,200 AD (Bellwood, 1987, 1997; Kirch, 2000; Diamond and Bellwood, 2003), with the various branches of Malayo-Polynesian derived along the voyage (Blust, 1996).

Archaeological evidence has been able to trace connections between the Chinese mainland, Taiwan, the Philippines and ISEA. In southwestern coastal Taiwan, an agricultural (rice and foxtail millet) economy, the Tapenkeng Neolithic culture, is present by at least 2,800 BC (Tsang, 2005; Tsang *et al.*, 2006). There have been numerous other sites discovered from 3,000 BC in eastern Taiwan (Hung, 2005), and the recovery of fine-grained ceramic evidence for the spread at about 2,200 BC of Neolithic material culture from Taiwan to the Batanes Islands (previously uninhabited) and northern Luzon (Bellwood and Dizon, 2005, 2008; Hung, 2005, 2008; Bellwood, 2011). Chicken, dog, and pig were domesticated in Asia and then brought into ISEA. The archaeological evidence, including red-slipped pottery

with specific rim forms and body shapes, pottery spindle whorls, stone bark cloth beaters, tanged or grooved stone adzes, Fengtian (eastern Taiwan) nephrite, Taiwan slate knives and projectile points, notched pebble net sinkers, suggest that these artefacts and domestic pigs and dogs, and possibly domesticated rice, were carried (not necessarily all together) at a single time or along a single route, namely, via Taiwan (Bellwood and Dizon, 2008; Bellwood, 2011).

The red-slipped plain ware pottery emerged by 2,200 BC in southern and eastern Taiwan (Bellwood and Dizon, 2008; Hung, 2008). One of the well-established red-slipped pottery assemblages came from Chaolaiqiao, on Shanyuan Bay, precisely dated by accelerator mass spectrometry (AMS)  $^{14}\text{C}$  to 2,200 BC. By 2,000 BC, this type of pottery tradition had spread to previously uninhabited Reranum and Torongan Caves on Itbayat Island, Batanes. The close similarities in pottery shared between Reranum and Chaolaiqiao possibly showed a direct migration from sites such as An Son in southern Vietnam to Itbayat occurred between 2,200 and 2,000 BC (Bellwood, 2011). A similar find of red-slipped plain ware (with small amounts of stamped and incised decoration) was also found at Bukit Tengkorak in Sabah around 1,300 BC, along with bark cloth beaters and trapezoidal cross-sectioned adzes paralleled in Batanes, Taiwan, and Fujian (Chia, 2003; Jiao, 2007). Besides, Talasea (Kutau/Bao) obsidian was found in Bukit Tengkorak possibly coming from the Bismarck Archipelago in Near Oceania, suggesting two-way human movement between 1,200-900 BC (Bellwood, 1989; Chia, 2003). These archaeological finds suggest that a red-slipped plain ware tradition has a clear Taiwan origin (Bellwood, 1997, 2011). There is no good evidence that showed cord-marked pottery, shell fishhooks, cut-shell beads, and shell adzes predate Malayo-Polynesian arrival in ISEA, although there is the widespread use of old shell for making artefacts (Bellwood, 1997).

Rice is detected from macrobotanical remains at archaeological sites in East Asia and is seldom found, except as inclusions in pottery, at purportedly early “Austronesian” or “Neolithic” sites in ISEA (Bellwood, 1997). Rice husks in pottery have been found in the cave of Gua Sireh and Lubang Angin in Sarawak and Bukit Tengkorak in Sabah, dating from ~2,200 BC onwards (Ipoi, 1993; Beavitt *et al.*, 1996), which are associated with paddle- or comb-impressed pottery with only rare red slip and no stamping. The rim forms and decoration in Gua Sireh is paralleled to the Middle Neolithic assemblages in southern Taiwan (Li, 1983) and Hong Kong (Meacham, 1978). Previously, the Gua Sireh assemblage was

reported to indicate a former Austro-Asiatic linguistic presence in Borneo (Bellwood, 2007:117, 236-238), but uncertainties arise when parallels of rice chaff-tempered pottery found in the southern Vietnam Neolithic sites, for example An Son in southern Vietnam (Bellwood, 2011). Apart from Gua Sireh, the above mentioned Borneo sites are far from fertile rice-growing terrain, and it is possible that the Neolithic and Iron Age burials in the upper Niah stratigraphy were not native people to the immediate area (Valentine *et al.*, 2008), where there was a continuing presence of hunter-gatherer population (Punan) in the Niah Cave until the Iban incursions in the 19<sup>th</sup> century (Barker, 2005).

Other early Austronesians-speakers went on to settle new islands very rapidly in terms of both archaeology and comparative linguistics (Pawley, 1999). The widely accepted “out of Taiwan” model suggests that the Austronesian-speaking populations of ISEA, Near Oceania, and Remote Oceania (including Polynesians) have a common origin among early Taiwanese agricultural groups who dispersed into ISEA ~4 ka, reaching Near Oceania ~3.5 ka (Bellwood, 2005a; Spriggs, 2007). The “Polynesian motif” represents a lineage of human mtDNA, B4a1a1a, which is restricted to Austronesian-speaking populations and is almost fixed in Polynesians. Based on the “out of Taiwan” model, these people are largely responsible for the Lapita culture complex, which includes finely decorated dentate-stamped pottery, obsidian tools, and shell ornaments that first appeared in the Bismarck Archipelago ~3.5 ka, spreading into Remote Oceania ~3 ka. Alternative models suggested maritime contacts between Southeast Asia and Near Oceania from the end of the Pleistocene ~12 ka (Solheim, 2006), or at least before the mid-Holocene, by ~6 ka (Terrell, 2004), that forms an interaction environment along a “voyaging corridor” between Near Oceania and ISEA (Irwin, 1992; Terrell and Welsch, 1997; Torrence and Swadling, 2008). Lastly, hybrid models suggest involvement of both incoming Austronesian speakers from ISEA and indigenous populations in the Bismarck Archipelago (Green *et al.*, 2008). A recent complete mtDNA genomes study by Soares *et al.* (2011) showed that the “Polynesian motif” most likely originated by >6 ka in and around the Bismarck Archipelago, and its immediate ancestor is estimated >8 ka and virtually restricted to Near Oceania. This suggests that the Polynesians have arrived from ISEA in Near Oceania much earlier than dispersal from either Taiwan or Indonesia 3-4 ka would predict. Soares *et al.* (2011) also reported evidence in minor lineages of more recent two-way maternal gene flow that reflects movements along a “voyaging corridor” between ISEA and Near Oceania, as previously proposed on archaeological find.

This work concludes a small-scale early to mid-Holocene Near Oceanic ancestry for the Polynesian peoples from ISEA, which transmitted Austronesian languages to the long-established Southeast Asian colonies in the Bismarcks carrying the Polynesian motif in the Lapita formative period, ~3.5 ka. Besides, the rapid movement reflects the people's dependence on the maritime and lowland agricultural resources, where the latter were reduced massively by the drowning of the most fertile alluvial and coastal locations as the sea attained its maximum mid-Holocene sea level (Bellwood *et al.*, 2008).

The “express-train” hypothesis for the colonisation of the Pacific by Austronesian-speaking peoples has been further supported by a parsimony analysis of a linguistic dataset of Austronesian languages studied by Gray and Jordan (2000). They studied the lexical similarities instead of its differences, and found that although the most parsimonious tree in the analysis was a close fit to the “express train” model, but many of the branches were not well supported. Besides, Gray and Jordan (2000) assumed a Taiwanese root in their analysis, where a Philippines root seems equally plausible too. Greenhill and Gray (2005) constructed a maximum likelihood (ML) tree with the same data with much better resolution and better supported branching structure. The ML tree also seemed to be consistent with the “express train” model but some of the branches, as before, were not as would be expected from the model. For example, the languages from North Borneo and Brunei appeared basally next to the Taiwanese languages.

Recent studies proposed a different Austronesian language tree in which the Western-Malayo-Polynesian and Central-Malayo-Polynesian do not exist as separate groups (Ross, 2008; Donohue and Denham, 2010). They did not find geographical structure within the Western-Malayo-Polynesian group's distribution. Other suggestions include that more branches are now radiating from Proto-Malayo-Polynesian node with no identifiable direction of dispersal, although their find points geographically to eastern Indonesia (Dyen, 1965). Donohue and Denham (2010) also questioned the validity of the Central-Malayo-Polynesian and Eastern-Malayo-Polynesian subgroupings since many of the innovations that have been proposed for each of these subgroups are present in languages in the Western-Malayo-Polynesian area as well as some area in north Taiwan (Donohue and Grimes, 2008).

## 1.9 *Orang Asli* in Peninsular Malaysia

In Peninsular Malaysia, the *Orang Asli* currently comprise 0.5% of the Malaysian population and have been categorised traditionally on the basis of language, culture, geographic location and anatomical traits (especially statures, hair type and skin colour). There are broadly three groups of *Orang Asli*, or literally “original people”; Semang, Senoi and Aboriginal Malays (Figure 1.9). Each group has its distinctive traditions, although authors like Benjamin (1985, 2002a, 2002b) and Rambo (1988) suggested that the Semang and Senoi shared a common origin purely as a result of local differentiation. However, many earlier authors claimed that the Semang and Senoi could have partially separate origins (Skeat and Blagden, 1906; Schebesta and Blagden, 1926).

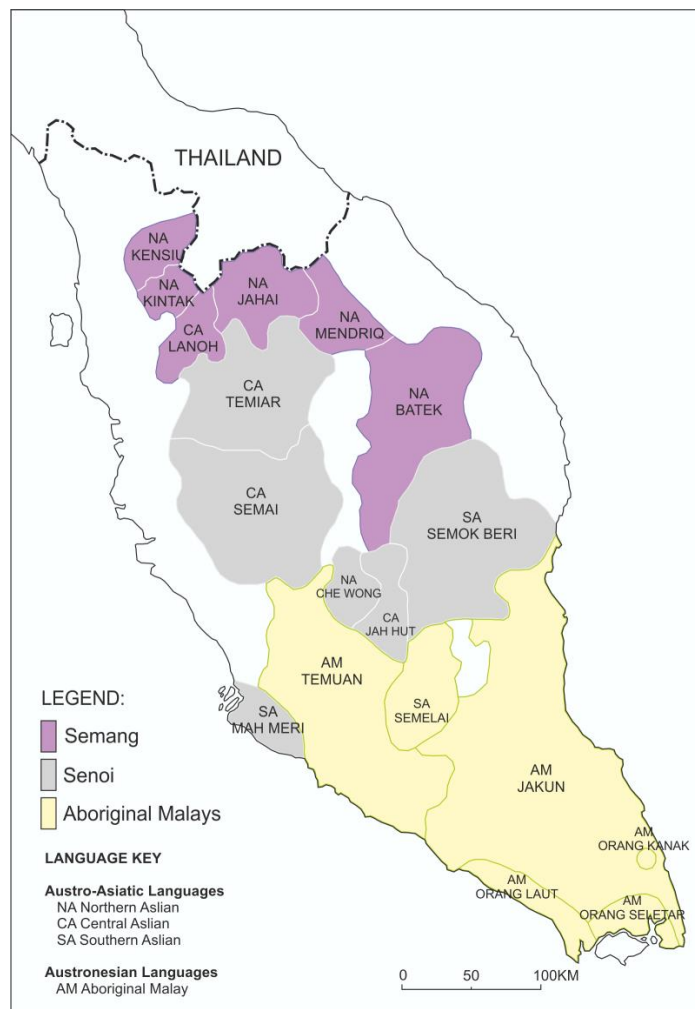


Figure 1.9 Map of Peninsular Malaysia showing the locations of *Orang Asli* groups sampled. Map adapted from Oppenheimer (1998).

There have been several views on the colonisation events that affected the *Orang Asli*. The most traditional view was the “layer-cake” theory of successive waves of arrival of Semang, Senoi, and Aboriginal Malays who settled in Peninsular Malaysia (Cole, 1945; Carey, 1976; Birdsell, 1993). The physical appearance of the Semang is closely related to the Philippine Aeta, Andaman Islanders, Melanesians, Tasmanians, and certain tropical Australian rainforest foragers. The Senoi arrived in the second wave of migration, probably originating from South Asia, and they were closely related to other small-bodied populations from Veddas of Sri Lanka, Toaleans of South Sulawesi, and mainland aborigines of Australia. Lastly, the arrival of Aboriginal Malays marked the first influx of Mongoloids into Peninsular Malaysia, in conjunction with the colonisation of the Indo-Malaysian Archipelago by the “Proto-Malays”.

Benjamin (1979, 1985, 1986) divides the Aslian languages into Northern, Central, and Southern subgroups. The Semang, such as Jahai, Mendriq and Batek, speak Northern Aslian languages, which is part of the Mon-Khmer language branch of the Austro-Asiatic family. This family is spoken widely in areas of northeast India, Burma, southernmost Thailand, Indochina, Peninsular Malaysia, the Nicobar Islands and the north of Sumatra (Ruhlen, 1991). Physically, they are short-statured and small with gracile facial features, dark-skinned with tightly curled hair; hence they are sometimes referred to as negrito (Bellwood, 1997). They are, or were until recently, hunter-gatherers living in small, nomadic groups in the lowland rainforests in the north, and were grouped with the similar negrito foragers of the Philippines and Andaman Islands. Linguistically, the Andamanese shows limited affinity with a few small isolates in Papua New Guinea and eastern Indonesia, and there is no widely accepted interpretation of the relationship of the Andamanese languages to the extant linguistic families of the South Asia region (Greenberg, 1971; Wurm and McElhanan, 1975; Ruhlen, 1991; Blevins, 2007).

The Senoi (e.g. Temiar and Semai) speak Central Aslian languages and live in the central higher altitudes where they practice sedentary swidden agriculture. They are taller in stature than the Semang, with lighter skin and wavy hair. The Aboriginal Malays (e.g., Semelai, Temuan and Jakun) live in the southern lowland forest of the peninsula. Some speak Southern Aslian languages, and others speak Austronesian languages including some Malay dialects. Physically they resemble Malay and are mostly short, light-skinned, and mostly

straight-haired and sometimes referred to as “Mongoloids” (see Bellwood, 1997:69-95). They subsist by collecting and trading forest produce, horticulture and fishing.

In Peninsular Malaysia, several archaeological sites radiocarbon dated back to the terminal Pleistocene and Holocene such as Kota Tamban (Zuraina and Tjia, 1988; Zuraina, 1990), Gua Gunung Runtuh (Mokhtar and Tjia, 1994) and Bukit Bunuh (Mokhtar, 2006), though no site in Peninsular Malaysia has produced positive evidence for habitation between 15,000 and 26,000 years BP (Zuraina *et al.*, 1994; Zuraina *et al.*, 1998; Mokhtar, 2006). This observation is probably due to scanty use of the present-day peninsular landmass at the height of the LGM, owing to the relocation of the coastally oriented population to lowlands now inundated by sea (Bellwood and Renfrew, 2003; Bulbeck, 2011). Sea levels between 26,000 and 43,000 years BP were higher than at the LGM, and evidence of habitation is found from the basal archaeological deposits of two rockshelters in southern Thailand, Lang Rongrien and Moh Khiew. However, these sites’ faunal remains are mutually exclusive which has been interpreted as indicating sporadic visits to Lang Rongrien by a coastally oriented population, unlike the more intensive habitation at Moh Khiew (Mudar and Anderson, 2007). After 15,000 years BP, sites with Hoabinhian-like cobble-stone-based industry begin to appear in Peninsular Malaysia at all altitudes, and the subsequent incorporation of Neolithic technology of polished stone tools and pottery (Bulbeck, 2004a, 2011). Sumatraliths and other pebble tools flaked on only one face tend to be a feature of Hoabinhian assemblages in the western half of the peninsula and in northern Sumatra, with no documented cases in southern Thailand and very rare instances in central or eastern Peninsular Malaysia (Adi, 2000). Sumatraliths’ restricted geographical and chronological distribution suggests a maritime interaction across the Melaka Strait as it underwent flooding during the early Holocene (Bulbeck, 2008, 2011).

Bellwood (1990) proposed that the core regions of equatorial and prehumid rainforest in Sundaland, particularly in the interiors of Borneo and Sumatra, were only sparsely occupied before the expansion of farmers along the coastlines and up the river valleys within the past 4,000-5,000 years. During the drier conditions of the last glaciation, the foragers were able to penetrate the more open forest conditions into the inlands, as shown at Tingkayu and Baturong (Hagop Bilo) of Sabah (Bellwood, 1988a) and Kota Tampan of Perak (Zuraina and Tjia, 1988). During the wetter and warmer Holocene the closed forest conditions would have



restricted them to coastal and riverine zones, except for the Hoabinhian site of interior Peninsular Malaysia.

Bellwood (1997) proposed two migratory patterns in Southeast Asia in relation to their language distribution. Southeast Asia's negrito would represent the relict descendants of the "Australo-Melanesian" foragers. Both Austro-Asiatic and Austronesian languages had their origins in South China. Austro-Asiatic was brought into SEA during the middle Holocene, followed by the Neolithic Austronesian-speaking farmers' expansion of Mongoloid appearance. He suggested that Austro-Asiatic spread southwards into MSEA using a mainland route, while Austronesian expanded from Taiwan to the Philippines, subsequently into Indonesia and Malaysia. At the beginning of the Neolithic period 2,000-1,200 BC, Bellwood (1993) noticed a dramatic cultural change in methods of burial and also the presence of a wide range of artefacts that have no precedent in the Hoabinhian culture. Therefore, he suggested that the Senoi are descended from Hoabinhian tribes who integrated with incoming Neolithic farmers, who also brought with them the Aslian languages currently spoken by most *Orang Asli* groups. Some of the Aboriginal Malays speak Austronesian languages, which are closely related to the modern Malay language and other Malay-Chamic languages in West Borneo, Sumatra and coastal Vietnam. This indicates they represent a separate migration via ISEA (Bellwood, 1993). There is little evidence of a cultural change in Peninsular Malaysia until the arrival of bronze and iron metal-working and new artefactual styles after 500 BC and Bellwood (1993) suggested that the Aboriginal Malays arrived from Sumatra sometime after 2,000 BP.

Conversely, in attempts to explain the differences between the *Orang Asli* groups, Rambo (1988) proposed a local differentiation model, where the Semang and Senoi developed from the same ancestral population but differentiated through adaptation to the distinct local ecology they came to occupy. On the other hand, Solheim (1980) suggested that the Semang are descended from Hoabinhians who lived on the coast, and that Senoi had an indigenous origin, descended from those who lived inland, with subsequent admixture from the newly-arrived Aboriginal Malays who brought the Neolithic culture. The first inhabitants in the interior mountains would have been subjected to a whole new environment and lifestyle, and adapted to swidden farming, causing them to diverge into a distinctive group – the Senoi. The Semang remained in the lowland forests having evolved from the earliest

populations and traded forest products for tools and food (Rambo, 1988; also Benjamin, 1985, 1997).

Bulbeck (1996, 2000) and Rayner and Bulbeck (2001) studied the physical anthropological aspects (including dental, cranial and postcranial) of the *Orang Asli*. A comparison of statures by Bulbeck (1996) found both Hoabinhian and Ban Kao people to be taller than Neolithic and modern *Orang Asli* populations. Based on the cranial evidence, Bulbeck also suggested that the Semang and Senoi had a common origin and began diverging in the early Holocene as a result of differing selection pressures, owing to the adoption of agriculture and gene flow between the Senoi and Malay. The ‘Mongoloid’ genetic contribution to the cranio-facial characteristics of the Senoi was proposed to originate from the expanding Malay (Bulbeck, 2000). Rayner and Bulbeck (2001) reported that the Semang and New Guinea populations shared an ancient dental morphology similar to that of Europeans and North Africans. Other Southeast Asian and Pacific populations, including the Aboriginal Malay, do not seem to possess this morphology. This indicated that the Aboriginal Malay migrated later into the peninsular, while Senoi were intermediate between the two *Orang Asli* groups, again suggesting that they are the intermediate between the Semang and Aboriginal Malay.

Bulbeck (2004a) offered an alternative model for the interaction and migration of *Orang Asli* populations based on foraging ranges and the linking trails through the peninsula’s forests (affecting their lifestyles). He suggested that the Hoabinhians foraged predominantly along well-cleared trails through the jungle, trails that tend to follow valleys and a degree of connections between the west and east sides of the Titiwangsa Mountains range, the backbone of Peninsular Malaysia (Bulbeck, 2003). However, the trails along the lowland stretches were disrupted by later incoming Aslian populations. Hence, Bulbeck (2004a) suggested that the Semang would have had to adapt to drastically reduced ranges in inaccessible regions, navigating the jungle with minimal reliance on the cleared trails. The Senoi may have arisen from the Aslian speakers who sought independence from the main commercial network up in the hinterland. Finally, the Aslian communities who adapted to living along the rivers and coasts, and thrived in growing international commerce may have provided the origins for the Aboriginal Malay who speaks South Aslian languages and dialects of Malay (Bulbeck, 2004a).

### 1.9.1 The “Negrito Hypothesis”

The negrito hypothesis depicts a shared phenotype of dark skin, short stature, and tight curly hair among various contemporary groups of hunter-gatherers in SEA, especially in the Andaman Islands, Malaysia, and the Philippines. The shared phenotype could be due to a common descent from a region-wide, pre-Neolithic substrate of humanity, or alternatively, convergent evolution. All Philippine negritos speak Austronesian languages, and all Malaysian negritos speak languages in the nuclear Mon-Khmer branch of Austro-Asiatic, and Andamanese remain distinct. The negrito hypothesis has been recently tested by multidisciplinary approaches and is reviewed here (Blust, 2013; Bulbeck, 2013; Chaubey and Endicott, 2013; Endicott, 2013; Heyer *et al.*, 2013; Jinam *et al.*, 2013; McAllister *et al.*, 2013; Stock, 2013). Blust (2013) examined the negrito populations through linguistic and cultural aspects and argued that the similarity of the names of the thunder god shared between the Malaysian and Filipino negrito populations suggested, in favour of the negrito hypothesis, a common cultural and linguistic past for these two populations at a time probably preceded the end of the Pleistocene, with the Andamanese possibly separating earlier.

Stock (2013) studied the stature of Andaman Islanders and Aeta foragers from the Philippines in relation to phenotypic variation among hunter-gatherer groups more globally, and he found no differences; and suggests that considerations of hypotheses of negrito origins need to go beyond stature as a defining phenotypic characteristic. Bulbeck (2013) agrees with Stock’s conclusion, and his osteology analyses also showed that the Semang shared some hints of Southwest Pacific affinities in cranial shape, dental morphology, and dental metrical “shape”. The Andamanese have been shown to resemble Africans in their craniometrics and South Asians in their dental morphology, while Philippine negritos resemble Mongoloid Southeast Asians in these respects and also in their dental metrics. The Andamanese and Semang (and Senoi) people have also been found to be more similar to each other, whereas Philippine negritos are dissimilar to both. Bulbeck (2013) reported that negritos are linguistically diverse and culturally heterogeneous and may differ according to mode of subsistence. Drift after initial-founding, admixture and environment also all have to be considered as possible mechanisms for regional differences in negrito morphology and stature.

Chaubey and Endicott (2013) examined the genome-wide autosomal SNP data for a shared history between the tribes of Little Andaman (Onge) and Great Andaman, and

between these two groups and the rest of South and Southeast Asia (both negrito and non-negrito groups). The Onge and Great Andamanese negritos are the closest genetic neighbours, the latter appear to have received a degree of relatively recent admixture from adjacent regional populations but also share a significant degree of genetic ancestry with Malaysian negrito groups. Chaubey and Endicott (2013) find the Onge are more closely related to Southeast Asians than they are to present-day South Asians. There are subsequent admixtures with neighbouring populations (both between negrito lineages and with non-negrito lineages), but found no evidence of a single ancestral population for the different groups traditionally defined as negritos in SEA. Jinam *et al.* (2013) analysed the admixture patterns and genetic differentiation in negrito groups (Jahai and Kensiu) and modern Malay from Peninsular Malaysia and Singapore, using genome-wide SNP data. They found possible traces of recent admixture in both the negrito groups with the Malay, which indicated that the admixture was as recent as one generation ago.

Heyer *et al.* (2013) identify two predominantly Philippine negrito mtDNA lineages, B4b1 and P9 (as well as P10), both are rarely found in any of the Southeast Asian, Southwest Pacific, or African populations, again indicating unique mtDNA haplogroups in the Philippine negritos. McAllister *et al.* (2013) analysed the mtDNA haplogroups by SNP hierarchical typing of short-statured Australian Aboriginal groups in Far North Queensland (FNQ) and Tasmania with those of other Australian Aboriginal populations and SEA negrito populations (Philippines Batek and Mamanwa, and mainland Southeast Asian Jahai, Mendriq, and Batak). The principle components analysis (PCA) and multidimensional scaling (MDS) results showed that the FNQ and Tasmanian mtDNA haplogroups cluster with those of other Australian Aboriginal populations and are only very distantly related to Southeast Asian negrito haplogroups. The result seems to coincide with finding by Delfin *et al.* (2011) who identified two Y chromosome haplogroups, C-RPS4Y and K-M9 that predominate among the Filipino negritos. These MSY (male-specific region of the Y chromosome) haplogroups represent founding lineages in the Asia-Pacific region that are also shared with indigenous Australians, and not found among the Filipino non-negrito populations. Hence, Delfin and colleagues conclude a possible divergence and subsequent gene flow between some Filipino negrito groups and indigenous Australians.

In overall, Blust's (2013) work seems to be the only one who argued linguistically for the negrito hypothesis in the negrito populations in Malaysia, Philippines and Andaman

Islands. Other fields of research like osteometry and genetics found no shared ancient ancestry between the three populations arguing against the negrito hypothesis. These three populations appear to share a similarity in physical appearance and mode of subsistence, if not more.

### **1.10 Modern Malay in Peninsular Malaysia**

The present-day modern Malay, AKA ‘Deutero-Malay’, in Peninsular Malaysia speak Malay language, which is a major language of the Austronesian family. It is believed that the Malay in Peninsular Malaysia consist of various sub-ethnic groups of different ancestral origins migrated from Indochina and the Indonesian archipelago centuries ago (Wheatley, 1961). It is suggested that the modern Malay in the west (*Melayu Minang*) and south (*Melayu Jawa* and *Melayu Bugis*) of Peninsular Malaysia are historically and culturally closer to the Indonesian populations compared to the Malay in the north-eastern regions (*Melayu Kelantan*) (Hatin *et al.*, 2011). They are also referred to as Deutero-Malay, the descendants of the Proto-Malay, who have had historical influences and genetic admixture from the Arab, Chinese, Indian, Javanese, Siamese, Sumatran and Thai traders (Comas *et al.*, 1998). An alternative theory by Fix (1995) suggested that the Deutero-Malay originated from southern China over the past 3-3.5 ka (after the migration of the Proto-Malay) who then intermarried with the Proto-Malay and traders of the ancient trade routes resulted in the diverse recent Deutero-Malay populations that is now known as the modern Malay (Fix, 1995; Bellwood, 1997).

### **1.11 Previous Phylogeographic Analysis**

The sequence and timing of processes leading to the settlement of Southeast Asia by modern humans remain extremely controversial. As mentioned earlier, for many years, the question was addressed primarily using indirect archaeological and linguistic evidence, leading to a consensus that the archipelago was largely re-settled within the last 6,000 years by Austronesian-speaking, rice-growing communities from South China/Taiwan (Diamond, 1988; Blust, 1996; Bellwood, 1997). These were assumed to have replaced and/or assimilated the hunter-gatherer populations that had formerly inhabited the region. So far, most of the archaeological and linguistic studies on prehistoric modern human activities in Southeast Asia have therefore focused mainly on the last 6,000 years, in particular on a Neolithic arrival

from south China and Taiwan. However, recent genetic evidence has challenged this consensus.

Saha *et al.* (1995), Gajra *et al.* (1994) and Gajra *et al.* (1997) studied allele frequency data from many genetic loci (the so-called ‘classical markers’) of the Semai Senoi and they argued from this that the population had undergone a long period of isolation. Saha *et al.* (1995) analysed polymorphisms in red blood cell enzymes and plasma proteins of 349 Semai Senoi. They found private alleles of both red cell glucose-6-phosphate dehydrogenase (G6PD) and 6-phosphogluconate dehydrogenase which they argued indicated that the Senoi may have been isolated genetically and have a long population history. Saha *et al.* (1995) also analysed the genetic distance by both cluster and principal components models on multiple alleles at up to 13 polymorphic loci, and found a close relationship between the Semai and the Khmer of Cambodia, as well as being more closely related to the Javanese than to their Malaysian neighbours – the Malay, Chinese, and Tamil Indians. The Senoi did not appear to have a real link with the Vedda of Sri Lanka. A similar conclusion was suggested by Gajra *et al.* (1994) and Gajra *et al.* (1997) who looked at several forms of lipoproteins in Semai from Betau, Pahang. They found one of the ancestral haplotypes of apolipoprotein E allele had risen to high frequency, indicating, they argued, a long population history for the *Orang Asli*.

In 1988, Harihara *et al.* studied the RFLP of mtDNA samples taken from the Philippine Aeta, Japanese Ainu, Japanese, Koreans and Vedda of Sri Lanka. They carried out analyses using maximum parsimony and genetic distance methods and found that the Japanese, Ainu, and Korean populations were closely related to each other, while Aeta was found to show a relatively close relationship to these three populations, and Vedda turn out to be quite different from the other four populations. Another early study on coastal and highland PNG mtDNA RFLP by Stoneking *et al.* (1990) highlighted the importance of extensive geographic sampling of a defined area that led to better understanding of the influence of geography on mtDNA variation in human populations. Stoneking *et al.* (1990) was one of the first reports to find the 9-bp COII/tRNA<sup>Lys</sup> intergenic deletion in mitogenome (Cann and Wilson, 1983; Wrischnik *et al.*, 1987) at nucleotide positions (np) 8281-8289 that characterised haplogroup B. The deletion occurs at 40% mtDNAs of the coastal populations, and is seen in Indonesia and fixed in Polynesia, while it is absent from the highland populations and Australia (Hertzberg *et al.*, 1989; Stoneking *et al.*, 1990; Redd *et al.*, 1995; Sykes *et al.*, 1995; Kayser *et al.*, 2006; Soares *et al.*, 2011). The results suggested the highland PNG populations have

more ancient and long-term isolation from one another and from coastal populations (Stoneking *et al.*, 1990).

Ballinger *et al.* (1992) examined the RFLP mtDNA of seven Asian populations including Malaysian Chinese, Malay and *Orang Asli* from Peninsular Malaysia, northern Borneo, Han Chinese from Taiwan, Vietnamese and South Korean. Their phylogenies were reconstructed with RFLP data and adopted a haplogroup annotation method then which was completely different from the current system. However, five of their Semai (Senoï) samples were found to form a population-specific branch in their phylogeny, a group which they called then as group I with haplotypes (73, 75, 78, 81, 82) (Ballinger *et al.*, 1992:142). Other haplogroups they identified included haplogroups “D”, “A”, “E” and “F” that were observed in nearly all of their Asian samples. Ballinger *et al.* (1992) found these *Orang Asli* populations of Peninsular Malaysia showed close affinities to the Austronesian-speaking Sabah Aborigines in northern Borneo and the people of coastal Papua New Guinea, implying that there was some degree of Austronesian admixture to the *Orang Asli* in Peninsular Malaysia.

Melton *et al.* (1995) studied the control region of the mtDNAs of the SEA and Polynesian populations, including 30 Semai Senoi samples from Peninsular Malaysia. The classic marker of 9-bp COII/tRNA<sup>Lys</sup> intergenic deletion was found in 37% of their Senoi samples, and they carried mutation at np 16217. All of these belonged to haplogroup B4a. They presented a neighbour-joining tree for the haplogroup B samples which indicated that the *Orang Asli* samples clustered closely with samples from the Philippines, East Indonesia, and Java than to Malay, or Barito-area Kalimantan, implying again a possible Austronesian influence.

The mtDNA and Y-chromosome data from throughout Mainland and Island Southeast Asia, including Peninsular Malaysia and Sumatra of Western Indonesia, both suggest that the picture is much less simple than the prevailing ‘Out of Taiwan’ model suggests (Capelli *et al.*, 2001; Macaulay *et al.*, 2005; Hill *et al.*, 2006, 2007; Soares *et al.*, 2008; Hunt *et al.*, submitted). Hill *et al.* (2006) showed using HVS-I data that haplogroups M21 and R21 have an ancient ancestry in Peninsular Malaysia estimated to the Upper Pleistocene (Macaulay *et al.*, 2005). Haplogroup M21a has the most common type present in the Semang and its derivatives are found in a minority of Malay, Aboriginal Malay and Borneo. This seems to indicate gene flow from Semang and Senoi in the north to Aboriginal Malays in the southern

part of Peninsular Malaysia and into Borneo (also see Adelaar, 2006). Interestingly, the negrito Andamanese exhibits mainly Indian subcontinental haplogroups M31 and M32 (Thangaraj *et al.*, 2005), while Omoto (1995) found the Philippine Aeta has its own unique classical blood markers that are not observed in other populations. The findings showed none of the Semang samples resemble M lineages of either the Andaman Islands or the Philippine Aeta, hence refuting the traditional notion of a specific shared ancestry at least on the maternal line between the negrito groups of the Andamanese, Peninsular Malaysia and the Philippines.

Haplogroups B and R9 are two familiar and widespread mtDNA haplogroups in Southeast Asia based on the control region analyses (Torroni *et al.*, 1994; Kivisild *et al.*, 2002; Yao and Zhang, 2002; Yao *et al.*, 2002a; Kong *et al.*, 2003b). The two main divisions of haplogroup B are B4 and B5; the majority of B haplogroups in ISEA falls within B4a. Haplogroup B4\* is most common in China (especially Yunnan province) and also common in Korea and Thailand (Hill, 2005), with B4a being most frequent among Taiwanese Aborigines and in the Philippines (Hill *et al.*, 2007) and dating to ~25.8 ka (Soares *et al.*, 2009). However, B4a1 is uncommon elsewhere and not found further west than Southeastern Borneo and Lombok, and is most common in the Moluccas (Hill, 2005). Hill *et al.* (2006) found a particular B5b type elevated to high frequency in the Batek Semang, probably by drift. This type appears to have been introduced fairly recently from ISEA because it is a derived type present only in the Batek, and the root type is found in Sumatra and eastern Indonesia, but not in Indochina (Hill C, Soares P, Mormina M, and Richards M, unpublished data).

Haplogroup R9 has two main branches, R9b and F (Kong *et al.*, 2003b), which diverged ~47 ka (Soares *et al.*, 2009). Haplogroup R9b was found at high frequency in the Aboriginal Malays. Complete mtDNA genome sequences of R9b in Hill *et al.* (2006) indicated a Pleistocene origin in Indochina, with early-Holocene dispersal southwards to Peninsular Malaysia and into ISEA. The finding, at least these R9b lineages, appeared to counter Bellwood's (1997) view of Aboriginal Malays arriving from ISEA in conjunction with the expansion of Austronesian speakers in the archipelago. However, this may not necessarily fully represent the lineages that probably did arrive in the Aboriginal Malays from ISEA.

Haplogroup F1a is common and widespread in SEA, where its subclade F1a1a is found largely in Temiar and Semai of Senoi. The root type is observed in Indonesia, Taiwan and



China (Hill *et al.*, 2006), and also in Thailand (Fucharoen *et al.*, 2001) and Vietnam by control region data. Haplogroup F1a1a has a mid-Holocene age in the Senoi, and probably dispersed from South China as early as 9,000 ya. In 2007, Hill *et al.* (2007) examined a total of 1026 control region mtDNA samples from across ISEA and Taiwan. One of their findings showed that haplogroup F1a and its two sister subclades F1b and F1c are found in MSEA, with F1b and F1c mainly restricted to South China, suggesting a possible origin for F1 and F1a in South China. This view is at least consistent with the Neolithic populations proposed by Bellwood (1993) where Peninsular Malaysia was inhabited by groups dispersed from central Thailand (associated with Ban Kao culture), and then intermarried with indigenous inhabitants to create the ancestors of the Senoi, bringing along Austro-Asiatic to Peninsular Malaysia at the same time.

Subclade N9a6a is found unevenly distributed in all 3 main *Orang Asli* groups, but is most diverse in the Aboriginal Malays, and is also shared with the Malay in Peninsular Malaysia and Indonesia. This derived lineage was dated to ~5,500 ( $\pm 2,600$ ) years by control region data (Hill *et al.*, 2006). Haplogroup N9a is rather widespread in mainland East Asia, and its subclade N9a6 is found typically in South China, Indochina, and Sumatra. Similar to that of haplogroup R9b, N9a has a deep ancestry in MSEA and a more recent expansion through Peninsular Malaysia into ISEA.

The HVS-I data showed haplogroups N21 and N22 are more diverse in the Aboriginal Malays compared to other *Orang Asli* groups, and may be associated with other Austronesian speaking populations in Taiwan, ISEA and Micronesia (Hill *et al.*, 2007). Haplogroup N21 is also found in some Peninsular Malaysia Malay and has much more diverse lineages in Indonesia than the Aboriginal Malays, pointing to an origin in island Southeast Asia and a recent dispersal into Peninsular Malaysia. N22 was previously found by Hill *et al.* (2006) in *Orang Asli* Temuan, which is rare but more diverse in Indonesia.

The Aboriginal Malays (Semelai) also have M7c3c (previously nominated as M7c1c in Trejaut *et al.* (2005) and Hill *et al.* (2006)) that may have arrived recently from offshore. M7c3c nested within the ancestral M7c\* which is more common and diverse in China. Haplogroup M7c3c dates to 8,300 ( $\pm 2,400$ ) years by HVS-I data (Hill *et al.*, 2006), which predates the Out of Taiwan event into Island Southeast Asia. Considering the lower standard error of the date, however, the signal is consistent with an expansion of Austronesian speakers, mariner-agriculturalists, or both, in the mid-Holocene, as proposed by Bellwood

(2004), from Taiwan into ISEA, followed by a small-scale dispersal into Peninsular Malaysia from Indonesia.

Hill *et al.* (2006) detected four colonisation events in the *Orang Asli* populations in Peninsular Malaysia; over 50 ka, ~10 ka, mid-Holocene and late Holocene. All three *Orang Asli* groups have roots dated to ~50 ka, and all have been affected by subsequent migrations to the peninsula. Although these dates bring to mind the traditional layer-cake theory (Carey, 1976), the theory of unchanged relicts of earlier population waves was unsupported by Hill *et al.* (2006). The phylogenetic differences reflect distinct ancestries to a greater degree than Rambo's (1998) local differentiation hypothesis would imply. At any rate, the role of local evolution for all three *Orang Asli* groups should be acknowledged for at least after the early Holocene, at the same time having several waves of immigration from the north affecting the Semang and Senoi, and from island Southeast Asia affecting the Aboriginal Malays.

Two recent genetic reconstructions (Bulbeck, 2011; Oppenheimer, 2011) include discussion of mtDNA evidence published from Hill *et al.* (2006) onwards. Firstly, they argued that distinctive Aboriginal Malay lineages appeared to be of broad Sunda, rather than solely ISEA, origin (Oppenheimer, 2011; Bulbeck, 2011; Hill *et al.*, 2006). While this observation grouped them with Semang and Senoi, their lineages differed uniquely from the latter two, who shared unique ancient local lineages of their own. Secondly, all three *Orang Asli* groups appeared to have Holocene admixture from further north in MSEA, likely Neolithic and from multiple sources and times, possibly consistent with linguistic suggestions of Neolithic Austro-Asiatic language shift among the *Orang Asli* groups. However, while the northern Neolithic influence is similar to Bellwood's Thai Ban Kao suggestion, the Da But culture in Vietnam was suggested as possibly a more likely source archaeologically (Bulbeck, 2011). Finally, modern Peninsular Malays appeared to derive more of their mtDNA ancestry from MSEA, and possibly South China, than from ISEA (Oppenheimer, 2011).

The mtDNA diversity of *Orang Asli* in Peninsular Malaysia is low because of their small populations and resulting genetic drift. The Aboriginal Malays are comparatively more diverse than any other *Orang Asli* groups, but still less diverse than lineages found in ISEA. Although the Archaeogenetics Research Group at the University of Leeds has studied much of Southeast Asia in some detail, there has been a weakness in the sampling coverage. Previous work was limited to 260 samples from eight *Orang Asli* groups: Batek, Jahai and Mendriq of Semang, Temiar and Semai of Senoi, and Temuan, Semelai and Jakun of

Aboriginal Malay. Moreover, they were mostly analysed only at the low resolution afforded by the mtDNA control region (Hill *et al.*, 2006). Only nine of these samples were completely sequenced, including a Malay sample (Macaulay *et al.*, 2005). I therefore decided to carry out complete mtDNA genome sequencing by increasing the present coverage of samples among the *Orang Asli* subgroups and equally importantly, the Malay populations in West Malaysia and Sumatra of Western Indonesia, as well as to expand potential source regions to include mainland and Island SEA samples in order to identify the lineages present in these populations, their origins and age estimation.

## 1.12 Objectives and Hypotheses

The overall objective of this study is to describe and examine complete mitochondrial DNA genome variation in Peninsular Malaysians and potential source populations in South Asia, Mainland Southeast Asia and Indonesia, in order to test the current hypothetical models of settlement history of Peninsular Malaysia and the surrounding regions.

There are several specific, testable hypothetic models of origins of the four broad groups today, Semang, Senoi, Aboriginal Malays and mainstream Malay who have settled in Peninsular Malaysia in prehistory in multiple events, with more recent ones including origins of mainstream Malay:

1. **The traditional layer-cake** structure in the three *Orang Asli* groups of Peninsular Malaysia (summarised in Carey, 1976), postulated three successive waves of arrival of three ancestral founding groups, settling in Peninsular Malaysia as ancestors of the modern indigenous groups in the following order: the Semang and Senoi ancestral groups both came separately from South Asia, while the Aboriginal Malays (aka ‘Proto-Malays’) arrived separately from Island Southeast Asia. The Malay or ‘Deutero-Malay’ arrived from ISEA as a fourth ‘layer’.
2. **The human ecology or local differentiation model** (Rambo, 1988; Benjamin, 1985, 1986) relates *Orang Asli* phenotypic variation to niche differentiation. This local-continuity model suggests the Semang and Senoi were both originally of coastal indigenous Peninsular origin, adapting physically to their respective different interior lifestyles, with inland Senoi subsequently admixing with the newly-arrived Neolithic Aboriginal Malays arriving from ISEA (see also Solheim, 1980).
3. **Bellwood’s (1997) three-wave model** involves an initial Pleistocene settlement by Australo-Melanesian negrito giving rise to the “Hoabinhians” and thence the Semang. These interbred with a wave of migrating Neolithic farmers, who also originated further north in MSEA, in the Ban Kao culture of Southern and Central Thailand, bringing with them the Aslian languages currently spoken by most *Orang Asli* groups, and giving rise to the Senoi. Aboriginal Malays subsequently arrived as Neolithic Austronesian-speaking farmers, migrating ultimately from southern China and Taiwan, but proximally from the south via ISEA, then north into Peninsular

Malaysia, followed a fourth wave (i.e. modern or Deutero-Malay) taking the same route 3-4,000 years ago.

From these models we can formulate a series of testable hypotheses concerning the settlement of the Peninsula.

- To what extent are the Semang the descendants of the earliest settlers, as all of the models broadly suggest? Would this correlate with the “Hoabinhians”? Have they admixed with other groups and if so what were the sources of the additional lineages? This can be tested by looking at the extent of the indigenous lineages not found elsewhere, as opposed to lineages shared with other populations. When did they arrive? This can be tested by looking at the time depth of the indigenous lineages. Where did they come from? The Southern Coastal Route suggests India, but before South Asian populations began to differentiate significantly.
- Are the Senoi from a source in India (traditional model), or indigenous to the Peninsula (Rambo, 1988; Benjamin, 1985, 1986), or the result of mid-Holocene Australo-Asiatic migration from Thailand (Ban Kao), perhaps with some assimilation of Semang, as Bellwood (1997) proposes – or from coastal Vietnam (Da But), as Bulbeck (2008) has suggested? This can be tested by comparing Senoi lineages with the Semang and other Southeast Asians in Thailand and Vietnam, and to South Asians.
- Are the Aboriginal Malays also indigenous to the Peninsula (Rambo, 1988; Benjamin, 1985, 1986) or the result of migrations from ISEA (Bellwood (1997) and the traditional model)?
- Are the Malay the result of migrations from ISEA as Bellwood (1997) and the traditional model propose or might they actually have some indigenous ancestry within the Peninsula? To what extent do they also have ancestry from India and China, as the historical evidence would suggest?

Some partial answers to these questions have already been proposed on the basis of mtDNA control-region variation (Hill *et al.*, 2006; Bulbeck, 2011; Oppenheimer, 2011), and these will also be considered, but this thesis proposes to provide more precise answers by using whole-mtDNA genomes.

## 2 Material and Methods

### 2.1 Samples

#### 2.1.1 Participants in this study

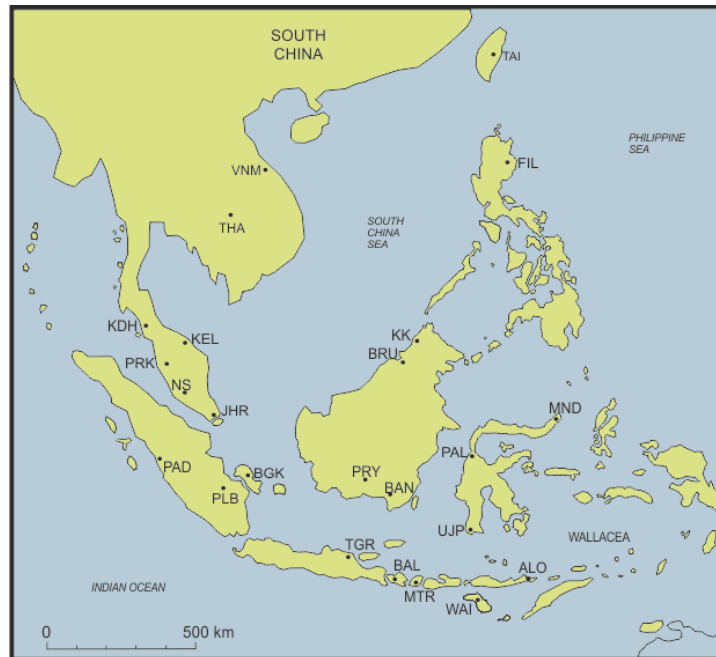
I collected cheek swabs from 85 maternally unrelated *Orang Asli* individuals from three Semang locations (after HVS-I sequencing, I reduced the number of samples from 144 to 85 to avoid duplication of haplotypes within families). The locations are situated in Lenggong and Pengkalan Hulu of Perak and Baling of Kedah respectively, all distributed along the Malaysian-Thailand border, and the samples were collected in January 2010. The samples include four subgroups of Semang maternal ancestry (nine Lanoh, 21 Kintak, 34 Kensiu and four Jahai), and two Senoi subgroups (one Semai and 16 Temiar). As this research involved human participants and biological samples (cheek swabs), I obtained the informed consent of each individual prior to sample collection. We also obtained the appropriate ethical clearance from the ethical board of the Faculty of Biological Sciences, University of Leeds. The cheek swabs were collected from the participants using OmniSwab (Fisher Scientific). I allowed the swab to air dry for 20 min, then replaced it in its plastic bag and sealed it up with label giving the sample code. I kept the swabs in the freezer at -20 °C.

Other samples include the previously analysed 260 *Orang Asli* (Hill *et al.*, 2006, 2007), and 297 modern Malay samples (Zafarina Zainuddin, personal communication and aliquots Dec 2011). The 297 modern Malay samples were collected from four regions in Peninsular Malaysia: 109 Northeast Peninsular Malay (10 Bachuk, 1 Tumpat, 42 Kota Bahru, 31 Rantau Panjang and 25 Machang of Kelantan), 98 Northwest Peninsular Malay (11 Yan and 26 Lembah Bujang of Kedah, 20 Kuala Kurau, 18 Parit Buntar and 23 Gopeng of Perak), 56 Southeast Peninsular Malay (20 Pontian, 6 Benut, 13 Semerah and 17 Muar of Johor), and 34 Southwest Peninsular Malay (22 Sri Menanti and 12 Lenggeng of Negeri Sembilan). The sampling locations in Peninsular Malaysia are shown in Figure 2.1. I re-sequenced the HVS-I region of the 297 Malay samples, from here I chose appropriate representatives from each haplogroup for complete mtDNA sequencing, which include 186 Malay samples and 40 *Orang Asli* (21 from Hill *et al.*, 2006 and 19 from my Semang samples).

K.C. Ang (personal communication) provided 91 HVS-I sequences of *Orang Asli* from Peninsular Malaysia and we have obtained permission to analyse these as well (the School of Environmental and Natural Resources Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia). Ang's 91 *Orang Asli* sequences consist of 18 subgroups (Table 2.1); they cover nps 16047-16567 in HVS-I with a size of ~520 bp each.

**Table 2.1** Distribution of three *Orang Asli* subgroups from Peninsular Malaysia including samples from Hill *et al.*, (2006), K.C. Ang and my *Orang Asli* samples.

OA in Peninsular Malaysia	Semang						Senoï						Aboriginal Malay/Proto-Malay						Total
	Batek	Jahai	Lanoh	Kensiu	Kintak	Mendriq	Chewong	Jah Hut	Mah Meri	Semai	Semok Beri	Temiar	Jakun	Kanak	Kuala	Seletar	Semelai	Temuan	
Hill <i>et al.</i> , 2006	29	51	/	/	/	32	/	/	/	1	/	51	2	/	/	/	61	33	260
K.C. Ang	5	5	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	91
K.K. Eng	/	4	9	34	21	/	/	/	/	1	/	16	/	/	/	/	/	/	85
	34	60	15	39	26	37	5	5	5	7	5	72	7	5	5	5	66	38	436



**Figure 2.1** Locations of the samples included in this study. Key: Kedah (KDH), Kelantan (KEL), Perak (PRK), Negeri Sembilan (NS) and Johor (JHR) in Peninsular Malaysia; Kota Kinabalu (KK), Brunei (BRU), Palangkaraya (PRY) and Banjarmasin (BAN) in Borneo; Padang (PAD), Palembang (PLB) and Bangka (BGK) in Sumatra; Tengger (TGR) in Jawa Timur; Bali (BAL), Mataram (MTR), Waingapu (WAI), Alor (ALO) in Nusa Tenggara Timur; Palu (PAL), Ujung Pandang (UJP) and Manado (MND) in Sulawesi; Thailand (THA); Vietnam (VNM); Philippines (FIL); Aboriginal Taiwanese (TAI).

### 2.1.2 Comparative published mtDNA complete sequences

I retrieved a total of 2206 published complete mtDNA genome sequences (obtained from GenBank, the 1000 Genomes Project (McVean *et al.*, 2012), and Archaeogenetics Research Group, Huddersfield) to build up the most complete phylogeny possible, for precise dating and to help make phylogeographic interpretations. The unpublished complete mtDNA genome locations from the Archaeogenetics Research Group, Huddersfield are shown in Figure 2.1. 1226 belonged to the haplogroup M, 978 belonged to N, including its major subclade haplogroup R, and one each of L3b1a1 and L4a1. The complete mtDNA genome data mainly came from Mainland and Island Southeast Asia (MSEA and ISEA) and surrounding regions, including East, Central, North and South Asia, Australasia, Polynesia and Melanesia. These sequences were published by the 1000 Genomes Project (McVean *et al.*, 2012), Andrews *et al.* (1999), Chandrasekar *et al.* (2009), Chaubey *et al.* (2008), Costa *et al.* (2009), Dubut *et al.* (2009), Dancause *et al.* (2009), Family Tree DNA (2010), Fornarino *et al.* (2009), Friedlaender *et al.* (2007), Gunnarsdóttir *et al.* (2011a), Gunnarsdóttir *et al.* (2011b), Hartmann *et al.* (2009), Hill *et al.* (2006, 2007), Ingman *et al.* (2000), Ingman and Gyllensten (2003), Jinam *et al.* (2012), Kong *et al.* (2003a, 2003b), Kong *et al.* (2006), Kong *et al.* (2011), Kumar *et al.* (2008), Li (2006), Loo *et al.* (2011), Macaulay *et al.* (2005), Mishmar *et al.* (2003), Nohira *et al.* (2010), Palanichamy *et al.* (2004), Peng *et al.* (2010, 2011a, 2011b), Pierson *et al.* (2006), Pradutkanchana *et al.* (2010), Rajkumar *et al.* (2005), Rani *et al.* (2010), Scholes *et al.* (2011), Soares *et al.* (2008), Soares, Rito and Richards (personal communication, 31/10/2012), Starikovskaya *et al.* (2005), Tabbada *et al.* (2010), Tanaka *et al.* (2004), Thangaraj *et al.* (2005), Thangaraj *et al.* (2006), Wang *et al.* (2011), Zhao *et al.* (2009) and Zheng *et al.* (2011)<sup>1</sup>.

The locations or regions of the complete sequences are indicated by three-letters codes in the phylogeography analysis. A complete list of the codes is shown in Appendix A.

---

<sup>1</sup> The haplogroup D complete sequences in Zheng *et al.* (2011) were included in the whole-mtDNA tree but these were not used in ML estimation due to time constraints.



## **2.2 Phenol-chloroform DNA extraction**

To each cheek swab, I added 44 µl of 10% sodium dodecyl sulphate (SDS), 200 µl RBC lysis buffer, and 5 µl of 50 µg/ml Proteinase K, to digest nucleic acid proteins and remove contaminants. The RBC lysis buffer consists of 0.32 M sucrose, 1% Triton × 100, 5 mM MgCl<sub>2</sub>·6H<sub>2</sub>O and 12 mM Tris-HCl pH 7.5. I incubated the mixture at 54 °C for 2 h and then 37 °C overnight.

Next day, 250 µl of 5M NaCl was added to the swab, which was left on ice for 40 min, followed by 30 min spin in a microcentrifuge at 13,500 rpm. I then added 500 µl of phenol:chloroform:isoamyl alcohol 25:24:1 (saturated with 10 mM Tris, pH 8.0, 1 mM EDTA, Sigma product number P3803), mixed and spun for 10 min. The upper aqueous phase was transferred to a new tube and the phenol:chloroform:isoamyl alcohol step repeated. After that, the upper aqueous phase was transferred to another new tube, I mixed in 150 µl of 7.5 M ammonium acetate and 1 ml of 100% ethanol (ice-cold) to the tube and left overnight at -20 °C.

On the third day, the tube was spun for 20 min at 13,500 rpm. Supernatant was decanted, and I washed the pellet with 500 µl ice-cold ethanol before spinning at 13,500 rpm for 10 min. After the supernatant was removed, the washing with ethanol was repeated again. The supernatant was removed after washing, and I left the pellet to air-dry. Lastly, I re-suspended the pellet in 100 µl of distilled water.

## **2.3 Whole-genome amplification**

Certain samples of low DNA concentration were whole-genome amplified by the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare). This amplification kit is capable to yield 4 µg to 7 µg from nanograms of DNA sample within 1.5 to 2 hours. The kit contains 0.9 ml of sample buffer, 0.9 ml of reaction buffer, 100 µl enzyme mix and 20 µl 10 ng/µl control DNA (lambda), which is enough for 100 reactions.

For each 1 µl (10 ng DNA) sample, I added 9 µl of sample buffer and denatured at 95 °C for 5 min. It was then immediately transferred onto ice (~4 °C) to prevent the reformation of double-stranded DNA. Next, I added 9 µl reaction buffer and 1 µl enzyme mix into the sample and started the incubation process at 37 °C for 2 hours, followed by the inactivation of

enzyme at 85 °C for 15 min. Finally, I diluted the amplified product with 180 µl distilled water.

## **2.4 Polymerase Chain Reaction (PCR) Amplification and Sequencing**

PCR is a powerful tool in molecular biology that can identify a specific sequence of DNA between two short oligonucleotide primers and amplify that sequence. The mtDNA amplification was carried out by means of a thermo-cycling reaction, i.e. cycles of denaturation, primer annealing and DNA extension. The final volume of each reaction was 35 µl, and it was carried out in a 96-well PCR plate. Each reaction contained 25.9 µl of deionized water, 7.1 µl of 5x buffer for GoTaq DNA polymerase (pH 8.5 with 7.5 mM MgCl<sub>2</sub>) from Promega™, 0.43 µl of 100 mM each deoxynucleoside triphosphate (dATP, dCTP, dGTP, and dTTP) from Bioline™, 0.18 µl of 100 pmol/µl each primer (Eurofins MWG Operon), 0.21 µl of GoTaq DNA Polymerase (Promega™) and 1 µl of DNA.

The samples were first amplified for HVS-I using primers set either 15873F (5'-TACTCAAATGGGCCTGTCCT-3') and 388R (5'-TGGTTAGGCTGGTGTAGGG-3') (Table 2.2), or 15256F (5'-AGACAGTCCCACCCTCACAC-3') and 131R (5'-ACAGATACTGCGACATAGGG-3'). Targeted samples were selected for complete mtDNA genome sequencing based on the HVS-I control-region variation, and if necessary as well the HVS-II. The complete mtDNA genome was amplified in 22 overlapping PCR fragments of around 900 base pairs (bp) each; using a set of 22-23 specifically designed nested primers with matching annealing temperatures (either Table 2.2 or Table 2.3). Each pair of primers was used individually in each PCR reaction, not multiplex. The temperature profile of the amplification reaction was 95 °C for 5 min in initial denaturation step, followed by 35 cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 70 s before the final extension step at 72 °C for 10 min.

**Table 2.2 22 pairs of nested primers used for complete mtDNA genome PCR amplification designed and optimised by Maria Pala (research fellow in the Archaeogenetics Research Group, Huddersfield).**

N°	Name	Primer Sequence (5'–3')	N°	Name	Primer Sequence (5'–3')
1	15873F	TACTCAAATGGGCCTGTCCT	12	7494F	CATGGCCTCCATGACTTTTT
	388R	TGGTTAGGCTGGTGTAGGG		8533R	TATTTGGAGGTGGGGATCAA
2	16413F	TGAAATCAATATCCCGCACA	13	8389F	ATGGCCCACCATAATTACCC
	727R	AGGGTGAACCTCACTGGAACG		9333R	GGAGCGTTATGGAGTGGAAG
3	449F	TTATTTTCCCCTCCCCTCC	14	9183F	CCTCTACCTGCACGACAACA
	1466R	GGCCCTGTTCAACTAAGCAC		10175R	GCACTCGTAAGGGGTGGAT
4	1331F	AAGGTGTAGCCCATGAGGTG	15	9815F	CCACGGACTTCACGTCATTA
	2342R	AGGCTTATGCGGAGGAGAAT		10858R	AATTAGGCTGTGGGTGGTTG
5	2007F	TGGTGATAGCTGGTTGTCCA	16	10609F	TAACCCTCAACACCCACTCC
	3169R	GGAAGGCGCTTTGTGAAGTA		11767R	GCGTTCGTAGTTTGAGTTTGC
6	2835F	CCAACCTCCGAGCAGTACAT	17	11402F	TGACTCCCTAAAGCCCATGT
	3894R	GGTTCGGTTGGTCTCTGCTA		12544R	TGGCTCAGTGTCAGTTCGAG
7	3587F	CCCTGGTCAACCTCAACCTA	18	12227F	CTAACTCATGCCCCCATGTC
	4526R	GATGAGTGTGCCTGCAAAGA		13299R	TTGGTTGATGCCGATTGTAA
8	4350F	CCATCCCTGAGAATCCAAAA	19	12913F	TCCAACCTCATGAGACCCACA
	5325R	TGATGGTGGCTATGATGGTG		14068R	AGGTGATGATGGAGGTGGAG
9	5162F	TCGCACCTGAAACAAGCTAA	20	13714F	GGAAGCCTATTCGCAGGATT
	6096R	TTACAAATGCATGGGCTGTG		14856R	AGGAGTGAGCCGAAGTTTCA
10	5888F	TACCTACCCCCACTGATGT	21	14478F	CAACCATCATTTCCCCCTAAA
	6959R	GCCACCTACGGTGAAAAGAA		15598R	GACGGATCGGAGAATTGTGT
11	6643F	TCCTACCAGGCTTCGGAATA	22	15195F	TATCCGCCATCCCATACATT
	7818R	AGGGCGATGAGGACTAGGAT		16439R	GCACTCTTGTGCGGGATATT

**Table 2.3 23 pairs of alternative nested primers used for complete mtDNA genome PCR amplification.**

N °	Name	Primer Sequence (5'–3')	N °	Name	Primer Sequence (5'–3')
1	15587F	CTCCGATCCGTCCCTAACAAA	13	8011F	AGTACTCCCGATTGAAGCCC
	155R	AATAGGATGAGGCAGGAATCAA		9218R	TTGGTGGGTCATTATGTGTTGT
2	16522F	TAAAGCCTAAATAGCCCACA	14	8459F	AACTACCACTTACCTCCCTC
	775R	AGGCATAGCGTTTTGAGCTG		9928R	AACCAGATCTACAAAATGCCAGC
3	584F	TAGCTTACCTCCTCAAAGCA	15	9742F	CAGAGTACTTCGAGTCTCCCTTC
	1612R	GCTACACTCTGGTTCGTCCAAG		10925R	AGGTTGGGGAACAGCTAAATAGC
4	1172F	CCTGGCGGTGCTTCATATCC	16	10279F	CCCTACCATGAGCCCTACAAAC
	2433R	GTGTTGGGTTGACAGTGAGGG		11472R	TTGAGAATGAGTGTGAGGCG
5	2182F	GCAGCCACCAATTAAGAAAGCG	17	11081F	ATAACATTCACAGCCACAGA
	3235R	CCTTAACAAACCCTGTTCTTGGG		12195R	GGTCGTAAGCCTCTGTTGTCAG
6	2815F	GGGCGACCTCGGAGCAGAAC	18	11654F	ACAGCCATTCTCATCCAAACCC
	4227R	ATGCTGGAGATTGTAATGGGT		12848R	GCTTGAATGGCTGCTGTGTTG
7	3598F	CTCAACCTAGGCCTCCTATT	19	12358F	ACCACCCTAACCTGACTTCC
	4552R	AAAAATCAGTGCGAGCTTAGC		13311R	TGCTAGGTGTGGTTGGTTGATG
8	4410F	CAGCTAAATAAGCTATCGGG	20	13134F	AGCAGAAAATAGCCCACTAA
	5483R	AGGTAGGAGTAGCGTGGAAGG		14371R	ATTGGTGCTGTGGGTGAAAGAG
9	4955F	CATAGCAGGCAGTTGAGGTGG	21	13930F	ATCACACACCGCACAAATCCC
	6345R	AGATGGTTAGGTCTACGGAGGC		14992R	AAGGTAGCGGATGATTCAGC
10	5871F	GCTTCACTCAGCCATTTACCT	22	14603F	GAAGGCTTAGAAGAAAACCC
	6831R	TGGTAGCGGAGGTGAAATATGC		15743R	GGAGGTCTGCGGCTAGGAG
11	6604F	CACCTATTCTGATTTTTCGG	23	15256F	AGACAGTCCCACCTCACAC
	7682R	GGAAAATGATTATGAGGGCG		15978R	AGCTTTGGGTGCTAATGGTG
12	7403F	ACCCTACCACACATTTCG			
	8560R	GGGCAATGAATGAAGCGAACAG			

To fill in the gaps of the complete genomes, which sometimes were not able to be amplified by the primers described above, I also used a third set of 32-pair nested primers, Table 2.4, published by Maca-Meyer *et al.* (2001). These primers produce shorter amplicons and were much more readily amplified. However, this PCR amplification mix was different from before and the volume for each reaction was 23.7 µl, which consisted of 18.1 µl deionized water, 5.0 µl of 5x Colorless GoTaq DNA polymerase buffer (pH 8.5 with 7.5 mM MgCl<sub>2</sub>) from Promega™, 0.3 µl of 100 mM deoxynucleoside triphosphate (dNTP) from Bioline™, 0.125 µl of 100 pmol/µl each primer (Eurofins MWG Operon), 0.15 µl of GoTaq DNA Polymerase (Promega™). Lastly, I added 1.3 µl of DNA to get a final reaction volume of 25 µl. I used the same temperature profile of the amplification reaction as described above.

**Table 2.4 32 pairs of alternative nested primers used for complete mtDNA genome PCR amplification (Maca-Meyer *et al.*, 2001).**

N°	Name	Primer Sequence (5'-3')	N°	Name	Primer Sequence (5'-3')
1	L16340	AGCCATTTACCGTACATAGCACA	17	L8299	ACCCCCTCTAGAGCCCACTG
	H408	TGTTAAAAGTGCATACCGCCA		H8861	GAGCGAAAGCCTATAATCACTG
2	L382	CAAAGAACCCTAACACCAGCC	18	L8799	CTCGGACTCCTGCCTCACTCA
	H945	GGGAGGGGGTGATCTAAAC		H9397	GTGGCCTTGGTATGTGCTTT
3	L923	GTCACACGATTAACCCAAGTCA	19	L9362	GGCCTACTAACCAACACACTA
	H1487	GTATACTTGAGGAGGGTGACGG		H9928	AACCACATCTACAAAATGCCAGT
4	L1466	GAGTGCTTAGTTGAACAGGGCC	20	L9886	TCCGCCAACTAATATTTCACTT
	H2053	TTAGAGGGTTCTGTGGGCAA		H10462	AATGAGGGGCATTTGGTAAA
5	L2025	GCCTGGTGATAGCTGGTTGTCC	21	L10403	AAAGGATTAGACTGAACCGAA
	H2591	GGAACAAGTGATTATGCTACCT		H10975	CCATGATTGTGAGGGGTAGG
6	L2559	CACCGCTGCCCAGTGACACAT	22	L10949	CTCCGACCCCTAACAACCC
	H3108	TCGTACAGGGAGGAATTTGAA		H11527	CAAGGAAGGGGTAGGCTATG
7	L3073	AAAGTCCTACGTGATCTGAGTTC	23	L11486	AAACTAGGCGGCTATGGTA
	H3670	GGCGTAGTTTGAGTTTGATGC		H12076	GGAGAATGGGGGATAGGTGT
8	L3644	GCCACCTCTAGCCTAGCCGT	24	L12028	GGCTCACTCACCACCACATT
	H4227	ATGCTGGAGATTGTAATGGGT		H12603	ACGAACAATGCTACAGGGATG
9	L4210	CCACTCACCCTAGCATTACTTA	25	L12572	ACAACCCAGCTCTCCCTAAG
	H4792	ACTCAGAAGTGAAAGGGGGCTA		H13124	ATTTTCTGCTAGGGGGTGGA
10	L4750	CCAATACTACCAATCAATACTC	26	L13088	AGCCCTACTCCACTCAAGCAC
	H5306	GGTGATGGTGGCTATGATGGTG		H13666	AGGGTGGGGTTATTTTCGTT
11	L5278	TGGGCCATTATCGAAGAATT	27	L13612	AAGCGCCTATAGCACTCGAA
	H5832	GACAGGGGTTAGGCCTCTTT		H14186	TGGTTGAACATTGTTTGTGG
12	L5781	AGCCCCGGCAGGTTTGAAGC	28	L14125	TCTTTCTTCTTCCCACTCATCC
	H6367	TGGCCCCTAAGATAGAGGAGA		H14685	CATTGGTCGTGGTTGTAGTCC
13	L6337	CCTGGAGCCTCCGTAGACCT	29	L14650	CCCCATTACTAAACCCACACTC
	H6899	GCACTGCAGCAGATCATTTTC		H15211	TTGAACTAGGTCTGTCCCAATG
14	L6869	CCGGCGTCAAAGTATTTAGC	30	L15162	CTCCCGTGAGGCCAAATATC
	H7406	GGGTTCTTCGAATGTGTGGTAG		H15720	GTCTGCGGCTAGGAGTCAAT
15	L7379	AGAAGAACCCTCCATAAACCTG	31	L15676	TCCCCATCCTCCATATATCC
	H7918	AGATTAGTCCGCCGTAGTCG		H16157	TGATGTGGATTGGGTTTTTATGTA
16	L7882	TCCCTCCCTTACCATCAAATCA	32	L15996	CTCCACCATTAGCACCCAAAGC
	H8345	TTTCACTGTAAAGAGGTGTTGG		H16401	TGATTTACGGAGGATGGTG

## 2.5 Gel electrophoresis

The PCR amplification products were visualized using gel electrophoresis. Electrophoresis is a technique that separates molecules (nucleic acids and proteins) according to their size and charge in a gel matrix run in an electric field. I prepared 2% agarose by adding 2 g of agarose in 100 ml of 0.5% TBE (Tris/Borate/EDTA) buffer with a drop of the fluorescent dye, 5 mM ethidium bromide. The dye intercalates between DNA strands and enables visualisation of the DNA bands under UV transillumination. I run the gel at a constant voltage of 80V for approximately 30 min. A molecular weight ladder of 0.13 µg/µl at 100 bp intervals (Promega™) was used to estimate the size of the PCR products and the concentration of the amplicons by comparing the intensity of the bands with those of the ladder.

## 2.6 DNA purification and sequencing

Our laboratory used two companies to purify and sequence the mtDNA fragments: GATC Biotech Ltd. (London) and Eurofins MWG Operon (Ebersberg, Germany). Both companies used the same purification and Sanger-based sequencing methods. The PCR products were first purified using the QIAquick purification kit (QIAGEN) or alternatively spin columns, and cycle sequencing on ABI 3730xl 96-capillary DNA Analyzer (AB Applied Biosystem, Foster City, CA, USA) using application of ABI Big Dye Terminator Kit associated with enzyme *TaqFS*.

I diluted the PCR products with sterilised water to obtain ~10 ng/µl of amplified DNA per sample before sending them for sequencing. Normally, the same amplification primers were used in the sequencing process. Alternatively, the following set of sequencing primers was used since they annealed downstream to the 3' end of the amplification primers and hence were more effective in sequencing (Table 2.5).

**Table 2.5 Primers for sequencing reactions, designed and optimised by Maria Pala. (Tm – annealing temperature)**

N°	Name	Sequence (5'-3')	Tm ( °C)	N°	Name	Sequence (5'-3')	Tm ( °C)
1	131R	ACAGATACTGCGACATAGGG	55.3	12	7614F	AAGACGCTACTTCCCCTATC	55.2
2	16521F	TAAAGCCTAAATAGCCCACA	55.3	13	8423F	CTATTCCTCATCACCCAACT	54.1
	739R	GTGGTGATTTAGAGGGTGAA	55.0	14	9213F	CACCAATCACATGCCTATC	54.7
3	614F	AATGTTTACGACGGGCTCAC	55.2	15	9922F	CCTGATACTGGCATTITGTAG	55.0
4	1402F	AAACTTAAGGGTCGAAGGTG	56.0	16	10689F	GGCCTAGCCCTACTAGTCTC	54.9
5	2176F	AAAGCAGCCACCAATTAAG	55.6	17	11452F	TGCCGCAGTACTCTTAAAAC	55.8
6	2897F	ATCCAATAACTTGACCAACG	55.1	18	12246F	CTAACAACATGGCTTTCTCA	54.0
7	3638F	TAGCCGTTTACTCAATCCTC	54.6	19	12973F	CTACTAGGCCTCCTCCTAGC	55.1
8	4410F	CAGCTAAATAAGCTATCGGG	54.6	20	13723F	TTCGCAGGATTTCTCATTAC	55.5
9	5191F	CACCCTTAATTCCATCCAC	55.2	21	14546F	ATAATAACACACCCGACCAC	54.8
10	5999F	TCTAAGCCTCCTTATTCGAG	54.1	22	15324F	CAAACTCCACCTCCTATTC	54.6
11	6643F	TCCTACCAGGCTTCGGAATA	59.7				

## 2.7 Data Manipulation

### 2.7.1 Variants scoring

The sequences were aligned against the revised Cambridge Reference Sequence (rCRS; Andrews *et al.*, 1999) using the Sequencher 5.0 software. Variants were recorded when the aligned positions were different from the rCRS. The data is collected in two formats, FASTA and as a table of variants in Excel. The Excel database contains the information of the samples, identified haplogroups, variants in the sequence, FASTA format, which make it easy to select samples for a specific network in .tor format and to build input files for PAML and BEAST.

A transition (A↔G or C↔T) was annotated by the position at which it differs from the rCRS, so that a transition at nucleotide position (np) 16189 is denoted '16189'; an extra nucleotide letter after the position number for transversion (A↔C, A↔T, T↔G, C↔G), for e.g. '16257A' is a mutation at np 16257 from cytosine (C) to adenine (A). An 'i' stands for an insertion and a 'd' for a deletion. I neglected any transversions to C and length polymorphisms around polycytosine (poly-C) tracts, which occur at nps 303-315 and 16184-16193 when np 16189 has a transition, because they are extremely frequent and are the result of heteroplasmy (Bendall and Sykes, 1995). A heteroplasmic position is annotated with the

position number followed by letter 'R'. The full version of Sequencher 5.0 can be used to export the aligned contig for one complete mtDNA genome as a single FASTA file. Alternately, the following steps were used to generate the individual FASTA file.

### **2.7.2 Error detection**

All sequences and traces were carefully read side by side. I checked every sequence against the existing HVS-I database to not only identify haplogroup status but also to detect sequencing errors. Error detection was carried out, which involves drawing networks of all the data on the basis of identification of errors outlined by Bandelt *et al.* (2001). Bandelt *et al.* (2001) identified five major types of errors in a taxonomy of artefacts: Type I base shift, II reference bias, III phantom mutations, IV base misscoring and V artefactual recombination. The procedure in error detection, any unclear signals in the traces, rare mutations like transitions or transversion at positions of low mutation rate, or potential heteroplasmic sites were re-sequenced. Any private mutations were taken as a potential error and I performed a second reading of the traces, or re-sequenced the sample if the chromatogram gave ambiguous reading.

### **2.7.3 v2nall**

I used an application software called v2nall, which was written by Dr. Vincent Macaulay of the University of Glasgow, to convert the variants/polymorphisms scored for each sample into FASTA/PHYLIP format. The input text file consists of variants of each sample per row excluding its sample ID. The variants were manually edited into five digits, for example, np 73 became 00073 and np 7819A was 07819A, and saved as a .tab file. I then executed the application in MSDOS and converted the variants into the consensus mtDNA sequence, which could be manually edited into individual FASTA/PHYLIP format. They were now readily usable by the bioinformatics tools.

### **2.7.4 Geneious**

Geneious 5.0 (Biomatters) is a program integrating bioinformatics and molecular biology tools for DNA, RNA and protein sequence alignment and analysis. Here, I used the basic feature to search and retrieve deposited nucleotide sequences associated with an article or data for an organism and genetic marker from NCBI GenBank. The search keywords I used consisted of "Homo sapiens complete mitochondri\* 16500:17000 [slen]" coupled with a country name (Southeast Asia and neighbouring country) or an article title. When the search



was completed, I selected and exported the sequences as a single text file or individual FASTA files.

### **2.7.5 Alignments by Clustal algorithm**

I used BioEdit and Sequencher 5.0, both Clustal-based programs, to align the sequences against the rCRS. However, BioEdit was not able to align the complete mtDNA sequences because of its length (16568 bp), and they had to be manually aligned by introducing gaps. In Sequencher 5.0, alignment of various length nucleotide sequences against the rCRS is done automatically and in a considerably shorter time. The aligned contig can only be viewed, not saved, in the demo version, but it could be saved or exported with the Sequencher USB key. For both programs, alignments were saved as a text, FASTA or PHYLIP file which was ready to be further used in mtDNASyn, Phylogenetic Analysis by Maximum Likelihood (PAML) and Bayesian Evolutionary Analysis by Sampling Trees (BEAST).

### **2.7.6 mtDNA-GeneSyn tool**

The mtDNA-GeneSyn tool is open-source software, developed for Windows-based platforms and implemented with the C++ language, which is available at: <http://www.ipatimup.pt/downloads/mtDNAGeneSyn.zip> (Pereira *et al.*, 2009). The tool identifies and classifies the mtDNA polymorphisms in a FASTA file containing aligned sequences against the rCRS. I clicked on the “Polymorphism” menu and chose “Import Aligned File” to import and process the aligned complete sequences. The output file lists the polymorphisms present in each sample, where they can be extracted and compiled into an Excel database. Under the same menu, the saved text file can be opened and converted into a Röhl data format (\*.rdf) format by choosing “Export to network file”, which is a binary matrix recognisable by the Network software. Alternatively, I used fm2net\_gui to prepare the binary matrix file.

### **2.7.7 fm2net\_gui**

fm2net\_gui is an executable application developed by Christopher Snell, in the Archaeogenetics group at the University of Leeds, to convert the mtDNA polymorphisms in a MS Excel file (\*.xls) into a Network input Torroni RFLP format (\*.tor), which we use for sequence variants (Figure 2.2). The input Excel file (\*.xls) consisted of three columns for sample code, variants and number of samples respectively. There are three output options, trimming to different calibrated lengths of HVS-I or converting the sequence as it is.

“Forster” deletes all transversions and any nucleotide position values above 16365 and below 16090; “Soares” deletes any values above 16400 and below 16051. I used the option “None” where no restrictions were applied, to convert the Excel file into a Network file (\*.tor).

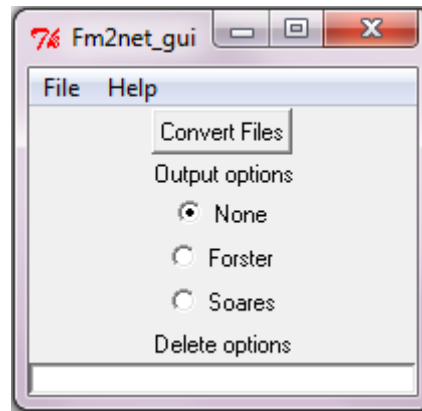


Figure 2.2 fm2net\_gui interface.

## 2.8 Phylogeographic Analysis

Phylogeographic analysis can be broadly defined as the analysis of the geographical distribution of the various mtDNA lineages within a phylogeny. A phylogenetic tree is used as the basic tool to understand the evolutionary processes of the mitochondrial data. There are four main methods for constructing phylogenies, namely Neighbour-Joining (NJ), Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian inference. Distance methods (NJ), which are non-character based, attempt to construct phylogenetic tree based on the estimated genetic distance for all pairs of sequences. However, they are considered as inferior to character-based methods because these algorithmic approaches strongly reduce the phylogenetic information of the sequences (to one value per sequence pair) (Van de Peer, 2009). MP infers a tree that has the least evolutionary changes to explain the observed data, while ML supposes the best hypothesis from a set of alternatives that maximises the likelihood of the outcome (Jobling *et al.*, 2004). Like MP and ML, Bayesian methods (such as can be carried out using the BEAST software) are character-based, but perform a probability analysis that allows inferences to be drawn from both the data and prior information. More information is described below.

### 2.8.1 Network 4.6

Traditional phylogeny constructing methods, such as distance methods and MP often fail to form nested sets of haplotypes for mtDNA data, and exhibit incompatibility between pairs of characters (Bandelt *et al.*, 1995) due to homoplasy (parallel or reverse mutations). For example, the number of equally parsimonious trees from just a set of small RFLP data (56 haplotypes) exceeded one billion (Excoffier and Smouse, 1994). Therefore, randomly choosing one single MP tree to represent the data can be both misleading and incorrect.

Network 4.6 implements the median-joining (MJ) and reduced-median (RM) methods (Bandelt *et al.*, 1995). Although they are parsimony method, they do not provide a single most parsimonious tree for a set of data, but summarise many by representing alternative evolutionary pathways with cycles or reticulations. The RM algorithm was chosen in this study over MJ, as the latter is a weaker and less reliable approach. During the RM calculation, the splits are based on the 0-1 combinations and each haplotype is represented by a 0-1 vector. The parsimonious network obtained has median vectors added to each triplet of sequences to represent unsampled data or extinct ancestral taxa (Bandelt *et al.*, 1995). RM networks highlight character conflicts in the form of reticulations, which can then be interpreted as homoplasy (e.g. high rates of homoplasy might lead to members of a single haplogroup being independently derived along different routes from the same ancestor), recombination, sequence error, or superimposed sequences (Bandelt *et al.*, 1995). The reticulations can be resolved by considering parsimony and frequency-based arguments in order to exhibit the most likely evolutionary routes through a network. Network can also highlight sequencing errors that manifests themselves in implausible network substructures.

The input file (\*.tor) generated by fm2net\_gui was imported into Network. The variants of each mtDNA sequence were then converted to a binary matrix with two states; 1 for presence, 0 for absence of a variant and saved as a \*.rdf file. Soares *et al.* (2009) identified all the major hot-spots present in the mtDNA genome based on the number of occurrence in the global mtDNA tree. The top ten hot-spots are sites from the control region: nps 16311, 16189, 16129, 16093, and 16362 in HVS-I and 152, 146, 195 and 150 in HVS-II. The sites can then be subjected to a weighing scheme when calculating a reduced-median network in order to eliminate some of the less plausible pathways. The default weight of 10 for each of the mutations is used, although it can be varied from 1 to 99, ideally not more than 15 because when too much weight was put on a character, it might obscure the true evolutionary

pathways and neglect equally acceptable ones. Fast sites in the HVS-I were down-weighted before calculating the reduced-median network. The weights of nps 16311, 16189, 16129, 16093 and 16362 were reduced from 10 to 3. The next fastest sites were reduced to 5; they were nps 16086, 16172, 16192, 16278, 16223, 16291, 16319 and 16390. This ranking is made according to Soares *et al.* (2009).

## 2.8.2 Phylogenetic trees

I manually drew the most parsimonious phylogenetic trees computed by the reduced-median networks for the complete mtDNA data on Microsoft Office Visio 2010. I inserted the mutations on the branches as described in 2.7.1. Insertions and deletions were shown only when they were phylogenetically informative. The PhyloTree.org website provides the basic phylogenetic tree framework of global human mtDNA and haplogroup nomenclature (van Oven and Kayser, 2009).

I then identified the type of mutation change for each mutation with an online Java-based program called MitoAnalyzer (Lee and Levin, 2000). The program evaluates single base-pair changes including mutations, insertions and deletions, and classifies them as tRNA, rRNA, control region and non-coding region mutations. The coding region is also subdivided into synonymous and non-synonymous mutations. An annotation is added to the end of the variants, e.g. s (synonymous), ns (non-synonymous), t (tRNA), r (rRNA), non-coding region mutation remains as it is. A reversion to the ancestral state was annotated by a symbol '@' and underlined, e.g. '@16189', and recurrent mutation in the tree was simply underlined. Variants in italics represent mutations that were associated with mitochondrial-related disorders. Fast mutations at nps 16182 and 16183, that were associated with 16189C, and np 16519 were excluded from the trees.

The sequence IDs and accession numbers were colour coded according to their geographic location (Figure 2.3). The *Orang Asli* were grouped by three main populations, dark green for Semang (Batek, Jahai, Lanoh, Kensiu, Kintak and Mendriq), mid-green for Senoi (Semai and Temiar), and light green for Aboriginal Malays (Jakun, Semelai, Temuan and Seletar). The Malay from Peninsular Malaysia were grouped by region: Northeast Peninsular Malay (Kelantan), Northwest Peninsular Malay (Kedah, and Perak), Southeast Peninsular Malay (Johor), and Southwest Peninsular Malay (Negeri Sembilan).





















	Semang		<i>East Asia</i>
	Senoi		China, Mongolia, South Korea & Xinjiang
	Aboriginal Malay		<i>West Pacific island chain</i>
	<i>Northwest Peninsular Malays</i>		Japan
	Perak & Kedah		Taiwan
	<i>Northeast Peninsular Malays</i>		<i>Oceania</i>
	Kelantan		Australia, Bougainville, Cook Is., Papua New Guinea, Samoa, Solomon Is., Tonga, & Vanuatu
	<i>Southwest Peninsular Malays</i>		
	Negeri Sembilan		
	<i>Southeast Peninsular Malays</i>		<i>South Asia</i>
	Johor		Andaman Islands, Bangladesh, India, Nepal & Pakistan
	<i>ISEA</i>		
	Indonesia – Java, Sumatra and Sulawesi		<i>Central Asia</i>
	South and North Borneo		Kyrgystan, Tibet & Uzbekistan
	Nicobars		
	Philippines		<i>North Asia</i>
			Russia & Siberia
	<i>MSEA</i>		America, Brazil, Colombia & Mexico
	Cambodia, Myanmar, Laos, Thailand, Vietnam		Madagascar

Figure 2.3 The colour codes for samples divided according to regional locations.

### 2.8.2.1 mtPhyl v4.015

The mtPhyl package is a useful free software tool for human mtDNA analysis and phylogeny reconstruction (Eltsov and Volodko, 2009). It has several helpful features: it compares an mtDNA sequence with the rCRS, and the variants can be exported into Excel format, or presented as a most-parsimonious phylogenetic tree (guided by the known phylogeny) in PowerPoint. Other tools include analysing mutation features, identifying mitochondrial haplogroups, calculating the coalescence time of nodes, and downloading complete human mtDNA sequences from GenBank. The aligned published complete sequences in 2.7.5 were saved as individual FASTA files which were imported into mtPhyl, since the program does not allow one to import aligned files of multiple sequences. With the advancement of sequencing technology today, the volume of sequence data is increasing rapidly, and mtPhyl helps to incorporate this massive amount of data into the phylogenetic trees. However, it has to be used with caution: the trees are approximate and have to be carefully checked and corrected, and the age estimates are not reliable.

## 2.9 Coalescence time estimation

I used three methods to estimate the coalescence age of the main clades: the rho ( $\rho$ ) statistic, Maximum Likelihood (ML) and Bayesian Inference (BI).

### 2.9.1 The rho ( $\rho$ ) statistic

The  $\rho$  statistic for coalescence time estimates was first described by Morral *et al.* (1994) and then Forster *et al.* (1996).  $\rho$  is the average number of sites differing between a set of sequences and a specified common ancestor. Given a sample of  $n$  sequences in a most-parsimonious tree with observed mutations and a specified root and  $m$  links, and take the number  $l_i$  as observed mutations along the  $i$ th link,  $\rho$  can be expressed as  $\rho = \frac{\sum_{i=1}^m n_i l_i}{n}$ .

The variance was originally defined as  $\sigma^2 = \frac{\rho}{n}$ , assuming all mutations in the estimate are independent, which is unlikely (and only occurs if the tree is perfectly starlike). A revised formula for the variance that takes into account the non-independence of lineages in non-starlike phylogenies is given as  $\sigma^2 = \frac{\sum_{i=1}^m n_i^2 l_i}{n^2}$  (Saillard *et al.*, 2000). The 95% confidence intervals were calculated as  $(\rho - 1.96\sigma; \rho + 1.96\sigma)$ . The posterior conversion in years, using the non-linear mutation rate corrected for purifying selection, was carried out using the calculator provided in Soares *et al.* (2009). The major advantage of  $\rho$  is that it is a simple, model-free estimate, and it can be applied to any clock.

### 2.9.2 Maximum Likelihood (ML)

ML is here used to calculate branch lengths in a pre-defined tree and was carried out with PAML software (Phylogenetic Analysis by Maximum Likelihood; Yang, 1997), stipulating a molecular clock in a given nucleotide substitution model. It is hence not used as a tree-building method since the RM network for mtDNA sequences is adequate for this.

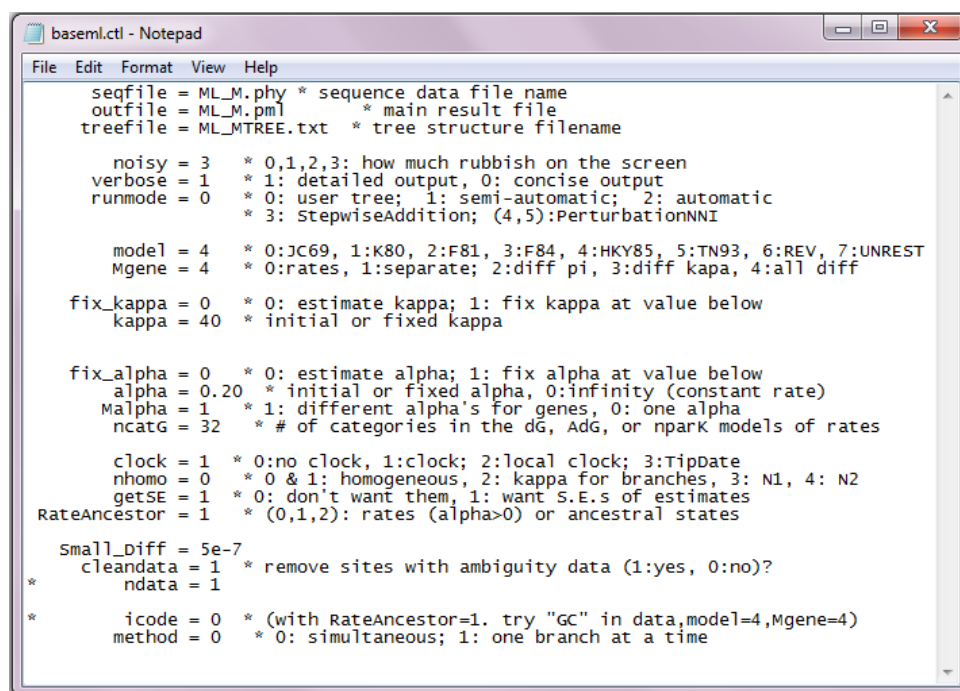
In the same folder, there are two input files for PAML analysis, the seqfile (.phy) and the treefile (.txt). To prepare the seqfile, sequences were aligned against the rCRS. Additionally, np 3107 (a spurious base in the original, erroneous CRS that was retained as a deletion to conserve the numbering in the rCRS) was changed from 'N' to 'T', and np 16519 was removed entirely since it was not considered in the calibration, resulting in a final length of 16568 bp. The seqfile begins with the number of samples, followed by the number of nucleotides, and ends with option character 'G'. The next row begins with another 'G'

followed by the number of partitions being used. For example, for 100 sequences using 2 partitions:

100 16568 G

G 2

Next, there is a ‘map’ of the partitions in the sequence corresponding to the HVS (partition 2) and the remainder of the molecule (partition 1). Finally, the newick treefile contains the shorthand text notation of the tree structure using nested parenthesis (which is recognisable by TreeView and FigTree tools). In the baseml.ctl file within the PAML package, the name of the seqfile, treefile and outfile are specified and saved (Figure 2.4).



**Figure 2.4** A screenshot of the baseml.ctl file opened in Notepad.

When all files have been prepared, the PAML analysis is executed by a double-click on baseml.exe. For example, the output file contains the length of each partition 16021, 547 (a total of 16568 bp), and their rates (1, 14.291754 respectively). The branch lengths  $b$  and standard errors (SE) are given for one whole gene (data set), for e.g.  $b$  0.000135, SE 0.000014. To obtain the ML genetic distance, I calculated the sum of partitions ( $0.000135 \times \text{Rate} \times \text{Length}$ );  $(0.000135 \times 1 \times 16021) + (0.000135 \times 14.291754 \times 547) = \text{final value}$ , to input into the calculator (Soares *et al.*, 2009). Similar calculations were applied for the SEs, which were finally converted to 95% confidence intervals.

### 2.9.3 Bayesian evolutionary analysis by sampling trees (BEAST)

Bayesian evolutionary analysis by sampling trees (BEAST) is a package that employs a Bayesian statistical framework for parameter estimation and hypothesis testing of evolutionary models from molecular sequence data. The core algorithm of BEAST uses Metropolis-Hastings Markov chain Monte Carlo (MCMC) sampling procedures to estimate a posterior distribution of effective population size through time directly from a sample of gene sequences, given any specified nucleotide-substitution model (Metropolis *et al.*, 1953; Hastings, 1970; Drummond *et al.*, 2005; Drummond and Rambaut, 2009). The Bayesian Skyline Plot (BSP) model uses MCMC and is able to co-estimate the evolutionary rate, substitution model parameters, phylogeny, and ancestral population dynamics within a single analysis, as well as to reconstruct demographic history under various expected scenarios (Drummond and Rambaut, 2007).

BEAST software package v1.7.4 contains BEAST, BEAUti, LogCombiner, TreeAnnotator, Figtree and Tracer (v1.5). BEAUti creates the input file .XML to be run in BEAST. The name of the complete mtDNA sequence was edited to carry details of its haplogroup, location code and sequence ID (e.g., B4a1a1a\_BOR\_B4274). The aligned sequences were saved as Nexus format (usually .nex or .nxs) by Sequencher and imported into BEAUti. Nexus format is widely used in phylogenetic programs (for e.g. PAUP and MrBayes) for storage and exchange of phylogenetic data such as store DNA and protein sequences, taxa distances, alignment scores and phylogenetic trees. Once imported into BEAUti, the sequences were grouped monophyletically, facilitated by the edited sample IDs, where each group has at least three or more sequences. The parameters are summarised in Table 2.6.

For BEAST analysis, I used a relaxed molecular clock (lognormal in distribution across branches and uncorrelated between them) and the HKY model (nst=2) of mutation with gamma-distributed rates. The HKY or HKY85 model (Hasegawa, Kishino and Yano, 1985) is considered as the extension of the Kimura80 and Felsenstein81 models, where it allows variable base frequencies and distinguishes between the rate of transitions and transversions (hence number of substitution types, nst=2). I applied a mutation rate of  $2.514 \times 10^{-8}$  variation/position/year, estimated for haplogroup U6 with four calibration points (Pereira *et al.*, 2010), as the molecular clock. I ran up to 450,000,000 iterations, with samples being drawn every 1,000 MCMC steps, and discarded burn-in of 10% or 45,000,000 steps. I



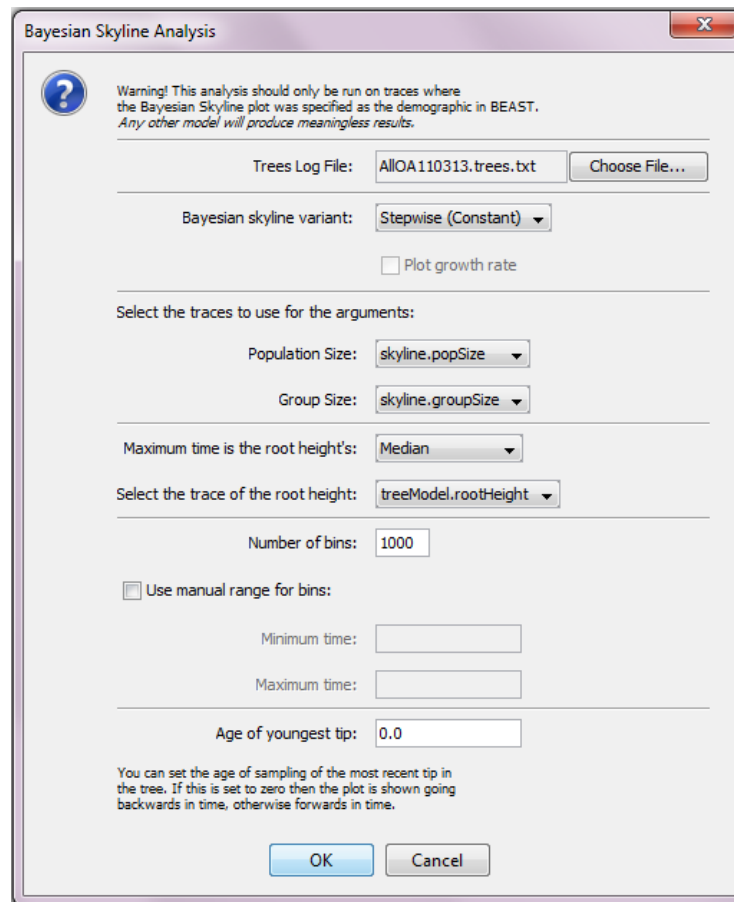
checked for convergence to the stationary distribution and sufficient sampling by inspection of the posterior samples.

**Table 2.6 The general settings for BEAUTi v1.7.4.**

<b>Tabs</b>	<b>BEAUTi v1.7.4 settings</b>	
Partitions	No change	
Taxa	Define the monophyletic groups (at least >3 sequences)	
Tips	No change	
Traits	No change	
Sites	Substitution Model:	HKY
	Base frequencies:	Estimated
	Site Heterogeneity Model:	Gamma
	Number of Gamma Categories:	10
	Partition into codon positions:	Off
Clocks	Name:	U6
	Model:	Lognormal relaxed clock (Uncorrelated)
	Estimate:	untick
	Rate:	2.51E-08 (Pereira <i>et al.</i> , 2010)
Trees	Tree Prior:	Coalescent: Bayesian Skyline
	Number of groups:	10
	Skyline Model:	Piecewise-constant
	Tree Model:	Choose Random starting tree
States	No change	
Priors	No change	
Operators	Tick Auto Optimize	
MCMC	Length of chain:	450,000,000
	Echo state to screen every:	1000
	Log parameters every:	1000
	File name stem:	abc.txt
	Log file name:	abc.log.txt
	Trees file name:	abc.trees.txt

### 2.9.3.1 Bayesian Skyline Plot (BSP)

Bayesian skyline plots were calculated by BEAST and visualised in Tracer. A BSP simulates the periods of major expansions (growth) as long as the data is sufficiently informative about the population. BEAST outputs two files, a .log.txt containing the posterior probability of the evolutionary parameters, and a .trees.txt with the trees generated from the sequences. TreeAnnotator can be used to summarise the information in the trees file produced by BEAST.



**Figure 2.5 Settings for Bayesian Skyline Plot in Tracer software.**

Tracer is a graphical tool for visualization and diagnostics of MCMC parameter output (Rambaut and Drummond, 2007). The outputs are the median age of the nodes monophyletically defined above, the corresponding mean age, the standard deviation, the 95% lower and upper highest posterior density interval (HPD) and the effective sample size. The Bayesian coalescence ages of the nodes are represented by the median age flanked by the corresponding 95% HPD intervals. The effective population size is inferred by piecewise reconstruction of the demographic history of the population at different points of time and space. I generated the BSPs for the *Orang Asli* and Malay populations in Tracer and exported data to Excel.

To compute the BSP, on Tracer's "Analysis" menu I selected "Bayesian Skyline Analysis" and located the trees file where it will be used to predict the BSP. Other settings are "Stepwise (Constant)" for "Bayesian skyline variant", "Select the traces to use for the arguments" that are default at "skyline.popSize" and "skyline.groupSize", "maximum time is the root height's" set to "median", and the "number of bins" to "1000" as shown in Figure 2.5.

### 3 Results and Discussion: Control Region and Haplogroup M

#### 3.1 Control-region variation

The primary results show, and analyse, variation in the complete mtDNA genomes of *Orang Asli* (OA, aboriginal peoples) and Malay (mainstream indigenous population) in Peninsular Malaysia in order to assist in the reconstruction of ancient population events. This process consists first of an overview examination of the populations by viewing and reconstructing phylogenetic networks of the short hypervariable segment 1 (HVS-I) sequences, followed by reconstructing high-resolution phylogenetic trees built using complete sequence mtDNA genomes. The Semang consist of six ethnic groups, Batek, Jahai, Mendriq, Lanoh, Kintak and Kensiu. Hill *et al.* (2006) analysed the first three of these groups in previous work and there were no representatives from the remaining three ethnic groups. Here I collected from the three further Semang groups (Lanoh (49), Kintak (47) and Kensiu (48)) in their northeastern interior locations of Peninsular Malaysia. To avoid duplication of haplotypes within families, these numbers were rationalised down, hence the lower numbers in Table 3.1. However, on reviewing individual maternal ancestral histories (two-generation) on the consent forms, a number of the participants' samples had mixed ancestry including both Semang and Senoi. This made it necessary to re-assign true maternal ancestry, which is reflected in Table 3.1 and detailed in its footnotes. On rationalisation and re-assignment of maternal ancestry, the three sampled Semang subgroups thus include 23 Kintak, 32 Kensiu, nine Lanoh, four Jahai, and two Senoi subgroups that include one Semai and 16 Temiar.

I characterised the HVS-I sequences corresponding with the results in the previous work. There were eight haplogroups found in these Semang and Senoi samples. The haplogroup frequencies of the samples are presented in Table 3.1. The reduced-median networks for haplogroup M and N were constructed with variants from the characterisation of HVS-I diversity. Results showed that the predominant clade among the Semang is M21a at 47.1% followed by R21 at 36.8%, reaching 56.0% in the Kensiu. Other haplogroups present at lower levels in Semang are R9b at 5.6%, M17a1a, N9a6a and B4c at 3.0% each, and M13b at 1.5%. In Senoi, there are 58.8% of F1a1a found in the Lanoh of Temiar ancestry, followed

by R21 at 17.6% and M21a at 11.8%. Haplogroups N9a6a and B4c are present at relatively low frequency among the Temiar and Semai at 5.9% each.

**Table 3.1 mtDNA haplogroup frequencies of 85 *Orang Asli* from Semang and Senoi populations<sup>2</sup>.**

Haplogroup	Semang				Total Semang (%)	Senoi		Total Senoi (%)
	Kintak	Kensiu	Lanoh	Jahai		Semai	Temiar	
M21a	15(0.65)	11(0.34)	5(0.56)	1(0.25)	32(47.1)	.	2(0.13)	2(11.8)
M13b	1(0.04)	.	.	.	1(1.5)	.	.	.
M17a1a	.	2(0.06)	.	.	2(3.0)	.	.	.
N9a6a	1(0.04)	1(0.03)	.	.	2(3.0)	.	1(0.06)	1(5.9)
R9b	2(0.09)	.	.	2(0.50)	4(5.6)	.	.	.
F1a1a	.	.	.	.	.	.	10(0.63)	10(58.8)
R21	2(0.09)	18(0.56)	4(0.44)	1(0.25)	25(36.8)	.	3(0.19)	3(17.6)
B4c	2(0.09)	.	.	.	2(3.0)	1(1.00)	.	1(5.9)
Total	23	32	9	4	68(100.0)	1	16	17(100.0)

91 HVS-I sequences representing all 18 ethnic groups of three *Orang Asli* groups (Semang, Senoi, and Aboriginal Malays) in Peninsular Malaysia were provided by K.C. Ang (personal communication, see Appendix B). There were five to six samples each from six Semang groups (Table 2.1 in Section 2.1.1). These data were found to provide valuable information because they include all *Orang Asli* ethnic groups and gives a good overall view of diversity. However, the sequences were problematic and impossible to check without their corresponding chromatograms. Many spurious transversions and ambiguous sites were observed in these HVS-I sequences that resulted in heavy reticulations in the reduced-median networks. Before I could phylogenetically examine the data, problematic sites needed to be identified through a series of networks in order to tease out probable sequencing artefacts.

Furthermore, 89 cytochrome B sequences (nps 14764-15174) were also provided with these 91 HVS-I sequences. However, only 84 samples appeared to have matching HVS-I and

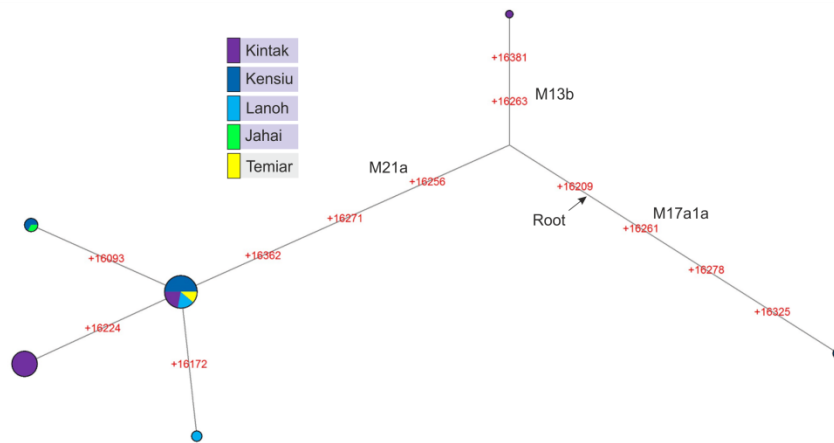
---

<sup>2</sup> In the haplotypes obtained from Kensiu villages, one M21a and two R21's are re-classified here as of Kintak maternal ancestry, while one R21 is reclassified as of Jahai maternal origin and one B4c is reclassified as of Semai, Senoi maternal ancestry. In the haplotypes obtained from Kintak villages, two M21a's, two M17a1a's and three R21's are here reclassified as Kensiu. In the haplotypes also obtained from Kintak villages, one M21a, and two R9b are here reclassified as of Jahai maternal origin. In the haplotypes obtained from the Lanoh, 2 M21a's, 1 N9a6a, 10 F1a1a's and 3 R21's are reclassified as of Temiar maternal origin. The latter major reclassification is consistent with the fact that the Lanoh are known to intermarry extensively with Temiar Senoi.

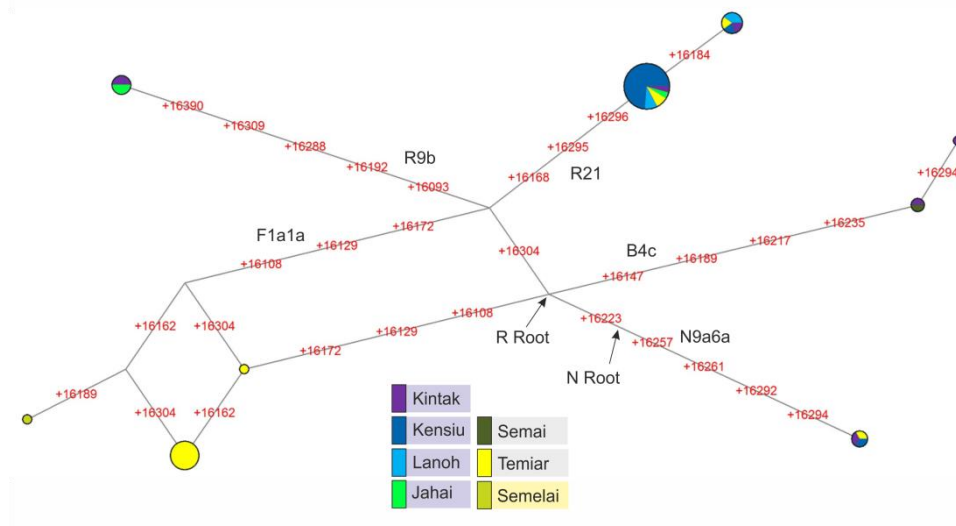
cytochrome B sequences. Cytochrome B provides crucial information in separating the haplogroup M samples from the N. The haplogroup M sample is defined by two transitions with respect to the rCRS in the cytochrome B sites at nps 14783 and 15043, which are absent in the N samples. However, some of the cytochrome B sequences of samples known to be haplogroup M from the HVS-I data had either one or other site, but not both, and 19 samples that were belonged to R21 on the basis of HVS-I had haplogroup M markers in their cytochrome B sequence. Therefore, the cytochrome B data was too unreliable to separate the sequences M from N, and not included in the network calculations. The haplogroup categorisation for these samples was thus done based on the control-region variants by comparing them with the existing data.

When the HVS-I sequences were aligned and scored against the rCRS, several uncommon mutations were detected throughout the sequences. These were nps 16059iA, 16192/16192iC/iCC/16192G, 16201N, 16516T and 16555T/iT. The score at np 16201N was likely a sequencing artefact as this ambiguous site can be resolved by resequencing. Insertions at nps 16059iA, 16192iC, 16192iCC and 16555T/iT were similar to Type III phantom mutations described by Bandelt *et al.* (2001), which could be caused by biochemical problems with the sequencing reaction. Transversion itself is rare and occurs at a much lower rate than transition. Apart from being an unlikely mutation such as the transversion from ‘G’ to ‘T’ at np 16516, they were not reported in the literature before. Besides, to have so many types of mutations at the same np 16192 only seems to suggest sequencing artefacts. They were therefore removed from the reduced-median network calculations as probable sequencing artefacts in order to generate more accurate networks.

Figure 3.1 shows the reduced-median network of HVS-I for 37 of my OA haplogroup M samples (Table 3.1), with the M root indicated on the network. The majority of these, 34 samples, belong to subhaplogroup M21a, which is defined here by control-region transitions at nps 16256 and 16271. There is one sample from M13b (previously M21b) that is defined by transitions at nps 16263 and 16381. The remaining two samples belonged to a novel haplogroup within M also found in previous work, which is now called M17a1a (Peng *et al.*, 2010; Tabbada *et al.*, 2010).



**Figure 3.1** The HVS-I Reduced-median network of haplogroup M for 34 samples. (Label boxes in light grey indicate Semang, and light blue for Senoi)



**Figure 3.2** The HVS-I Reduced-median network for haplogroup N and R for 48 samples. One Semelai of Aboriginal Malay sample (ORA; Hill *et al.*, 2006) from haplogroup F1a1a with a transition at np 16304 was included in the network to define the R root. (Label boxes in light blue indicate Senoi, light grey for Semang, and light yellow for Aboriginal Malay)

Figure 3.2 shows the HVS-I reduced-median network of haplogroups N and R for 48 samples of my *Orang Asli* data (Table 3.1), with the N and R roots indicated on the network. Haplogroup B is one of the most common haplogroups in ISEA, consisting of clades B4 and B5. Haplogroup B is defined, albeit inadequately, by a 9-bp deletion in the coding region and a fast transition at np 16189 in HVS-I (Soares *et al.*, 2007). Haplogroup B4c, seen in Semang, Kintak and Senoi, Semai had transitions at nps 16147, 16189, 16217 and 16235. The same variants were previously reported in Sumatran samples by Hill *et al.* (2006) and the root type of B4c is seen in Medan, Bangka and Palembang of Sumatra.

Figure 3.3 shows the HVS-I reduced-median network of all observed haplogroups scored from the 18 *Orang Asli* subgroups (KC Ang's data). Instantly, some transversions were observed to wrongly define certain haplogroups like those in clades R21 and M. In clade R21, transversions at nps 16197A and 16192G were postulating more new branches and even went on to define several samples. The same situation happened to some M samples where nps 16214A and 16100T were behaving in a similar way, where np 16214A was creating a new tip of the network as a private mutation. Bandelt *et al.* (2001) mentioned this is how artefacts normally manifest in those seeming private mutations. Apart from their absence in previous works, these transversions occurred frequently in the data suggesting they were sequencing artefacts.

Figure 3.4 shows the HVS-I corrected reduced-median network for all haplogroups (K.C. Ang's data) after the probable artefactual transversion sites 16100T, 16192G, 16197A and 16214A were removed. The network now better distinguishes the haplogroups observed in the 18 *Orang Asli* subgroups (K.C. Ang), although some reticulations are still present. Not all haplogroups scored for the same haplogroup shared the same variants. By reading the sequences against their chromatograms, which unfortunately were not available to us, ambiguous sites could have been checked by examining the signals of the traces. Good clear signals/peaks provide sequences of higher level of confidence. In order to better examine the lineages, I calculated reduced-median networks for haplogroups M and N separately.

Figure 3.5 shows the HVS-I reduced-median network for haplogroup M (K.C. Ang's data) after the removal of problematic sites, such as nps 16100T, 16214A, 16192iC from the calculation. Apart from the predominant haplogroup M21a, M7c is found in the Semelai of Aboriginal Malay, similarly reported by Hill *et al.* (2006), where it is also found in Malay, and other Austronesian speaking populations in Taiwan, ISEA and Micronesia. The ancestor M7c\* is the most diverse and common haplogroup in South China, and is believed to have dispersed into ISEA and then into Peninsular Malaysia. The age of haplogroup 'M7c1c' (currently it is M7c3c) was estimated at ~8,300 ( $\pm 2,400$ ) years ago in the ancestry of the Aboriginal Malays, suggested an origin from ISEA and Indonesia into Peninsular Malaysia (Hill *et al.*, 2006).





In Figure 3.6, the HVS-I reduced-median network for haplogroups N and R of 56 samples (K.C. Ang's data) is shown. Haplogroup R21 is still present in a tangle of reticulations, and although it is free of obvious sequencing artefacts, there were probably still some errors. The reticulation is due to different clusters sharing one to two variants with one another, hence when it is presented in a network, the nodes are unavoidably connected to each other in reticulations.

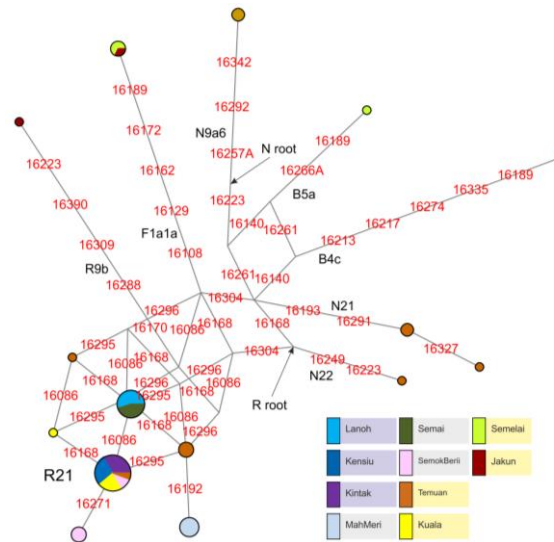


Figure 3.6 The HVS-I reduced-median network for haplogroups N and R of K.C. Ang's 56 samples. The roots N and R are indicated in the network.

Table 3.2 mtDNA haplogroup distribution of 18 *Orang Asli* subgroups (K.C. Ang).

Haplogroup	Semang					Senoi					Aboriginal Malay/Proto-Malay					Total(%)			
	Batek	Jahai	Lanoh	Kensiu	Kintak	Mendriq	CheWong	Jah Hut	Mah Meri	Semai	Semok Beri	Temiar	Jakun	Kanak	Kuala		Seletar	Semelai	Temuan
M*	.	.	.	.	.	.	.	.	.	.	.	.	.	4	5	.	.	.	9(9.9)
M21a	5	.	.	.	.	4	5	5	.	.	.	.	.	.	.	.	.	1	20(22.0)
M13b	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1(1.1)
M7c3c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1(1.1)
E	.	.	.	.	.	.	.	.	.	.	.	.	2	1	.	3	.	.	6(6.6)
N21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	3(3.3)
N22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1(1.1)
N9a6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	2(2.2)
R9b	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	1(1.1)
R21	.	5	6	5	5	.	.	.	5	5	5	5	.	.	.	.	.	1	42(46.2)
F1a1a	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	2	.	3(3.3)
B4c	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1(1.1)
B5a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1(1.1)
No. of samples	5	5	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	91(100.0)

Table 3.2 shows the mtDNA haplogroup distribution of 18 *Orang Asli* subgroups (K.C. Ang's data). The Semang and Senoi show considerably lower diversity, where the clades are predominantly haplogroups M21a and R21, and one individual of M13b was sampled from

Mendriq, Semang. In the Aboriginal Malay, the haplogroup diversity is much higher than the northern groups, as noted before by Hill *et al.* (2006). The limited number of sequence types and high levels of haplogroup sharing suggest that these *Orang Asli* populations lost diversity through drift.

Table 3.3 shows the combined data of 436 *Orang Asli* samples and haplogroups from 260 samples in Hill *et al.* (2006), 91 K.C. Ang and 85 samples from this study. Again, M21a remains the predominant clade at 25% among all the *Orang Asli* samples. It is seen in all six Semang subgroups, as well as in the neighbouring Senoi subgroups CheWong and Jah Hut, and in the Aboriginal Malay (Temuan and Semelai). R21 is present in 19.5% of the samples, especially in all Semang subgroups. It is also observed among Senoi, Temiar (31%), who are situated immediately to the south and west of the Semang in northern Peninsular Malaysia, and Semelai and Jakun of Aboriginal Malay that are next to each other in the south. The third common haplogroup that is present in all three *Orang Asli* subgroups is F1a1a at overall 18.3%, but particularly the Senoi. R9b is seen in 6.7% of the samples in the Semang and Aboriginal Malay. B5b is seen only in the Batek and Mendriq of Semang at 38% and 5% respectively (3.4% of all the samples). N9a6 (5.0%) and M13b (2.8%) are present in the *Orang Asli* subgroups in a rather similar distribution and frequency pattern, where they are found in almost half of Semang and Aboriginal Malay, and Temiar, Senoi. The genetic makeup of Aboriginal Malay subgroups, in particular Semelai and Temuan, are more diverse compared to Semang and Senoi. Haplogroups found only in Aboriginal Malay include N21 (6.2%), M7c3c (2.1%), E (1.4%), M22 (1.4%), N22 (1.1%), M21c (0.5%), and M7c1a (0.2%).

**Table 3.3 Combined *Orang Asli* subgroups and haplogroups for mtDNA HVS-I of 260 samples in Hill *et al.* (2006), 91 K.C. Ang (personal communication) and 85 (this study).**

Haplogroup	Semang						Senoi						Aboriginal Malay						Total(%)
	Batek	Jahai	Lanoh	Kensiu	Kintak	Mendriq	CheWong	Jah Hut	Mah Meri	Semai	Semok Beri	Temiar	Jakun	Kanak	Kuala	Seletar	Semelai	Temuan	
M*	.	.	.	.	.	.	.	.	.	.	.	2(0.03)	1(0.14)	4(0.80)	5(1.00)	.	1(0.02)	1(0.03)	14(3.2)
M21a	19(0.56)	9(0.15)	5(0.33)	11(0.30)	15(0.54)	31(0.84)	5(1.00)	5(1.00)	.	.	.	5(0.07)	.	.	.	.	2(0.03)	2(0.05)	109(25.0)
M21c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2(0.03)	.	2(0.5)
M13b	.	2(0.03)	.	.	1(0.04)	2(0.05)	.	.	.	.	.	1(0.01)	.	.	.	.	4(0.06)	2(0.05)	12(2.8)
M22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	6(0.16)	6(1.4)
M17a1a	.	.	.	2(0.05)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2(0.5)
M7c1a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1(0.02)	.	1(0.2)
M7c3c <sup>3</sup>	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	9(0.14)	.	9(2.1)
E	.	.	.	.	.	.	.	.	.	.	.	.	2(0.29)	1(0.20)	.	3(0.60)	.	.	6(1.4)
N21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	20(0.30)	7(0.18)	27(6.2)
N22	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5(0.13)	5(1.1)
N9a6	.	9(0.15)	.	1(0.03)	1(0.04)	.	.	.	.	.	.	4(0.06)	.	.	.	2(0.40)	1(0.02)	4(0.11)	22(5.0)
R9b	.	2(0.03)	.	.	2(0.07)	.	.	.	.	.	.	.	1(0.14)	.	.	.	17(0.26)	7(0.18)	29(6.7)
F1a1a	.	5(0.08)	6(0.40)	5(0.14)	5(0.18)	.	.	.	5(1.00)	6(0.86)	5(1.00)	37(0.51)	1(0.14)	.	.	.	4(0.06)	1(0.03)	80(18.3)
R21	1(0.03)	33(0.55)	4(0.27)	18(0.49)	2(0.07)	2(0.05)	.	.	.	.	.	22(0.31)	1(0.14)	.	.	.	2(0.03)	.	85(19.5)
B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3(0.08)	3(0.7)
B4a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2(0.03)	.	2(0.5)
B4c	.	.	.	.	2(0.07)	.	.	.	.	1(0.14)	.	.	1(0.14)	.	.	.	.	.	4(0.9)
B5a	1(0.03)	.	.	.	.	.	.	.	.	.	.	1(0.01)	.	.	.	.	1(0.02)	.	3(0.7)
B5b	13(0.38)	.	.	.	.	2(0.05)	.	.	.	.	.	.	.	.	.	.	.	.	15(3.4)
Total	34(1.00)	60(1.00)	15(1.00)	37(1.00)	28(1.00)	37(1.00)	5(1.00)	5(1.00)	5(1.00)	7(1.00)	5(1.00)	72(1.00)	7(1.00)	5(1.00)	5(1.00)	5(1.00)	66(1.00)	38(1.00)	436(100.0)

<sup>3</sup> M7c3c was previously named M7c1c (Hill *et al.*, 2006).

**Table 3.4 Distribution of the modern Malay samples grouped according to sample regions and haplogroups. The four regions in Peninsular Malaysia are Northeast Peninsular Malay (NEM), Northwest Peninsular Malay (NWM), Southeast Peninsular Malay (SEM), and Southwest Peninsular Malay (SWM).**

No.	Haplogroup	NEM	NWM	SEM	SWM	Total	%
1	A	0	0	1	0	1	0.3
2	B4a	3	5	1	2	11	3.7
3	B4a1a	4	3	1	0	8	2.7
4	B4b1	0	1	0	0	1	0.3
5	B4c1b2	3	4	4	10	21	7.1
6	B4c2	1	4	0	0	5	1.7
7	B5a	9	4	4	0	17	5.7
8	B5b	0	3	1	0	4	1.3
9	B6a1a	0	1	1	3	5	1.7
10	C7a	1	1	0	0	2	0.7
11	D4a3	1	0	0	0	1	0.3
12	D5b	0	0	1	0	1	0.3
13	E1a1a	2	1	7	0	10	3.4
14	E1a2	0	0	3	0	3	1.0
15	E1b	3	3	1	0	7	2.4
16	E2a	0	0	4	0	4	1.3
17	F1a1	2	2	0	0	4	1.3
18	F1a1a	8	7	2	0	17	5.7
19	F1a3	2	1	0	0	3	1.0
20	F1a4	1	2	0	0	3	1.0
21	F1f	6	2	0	1	9	3.0
22	F3a	1	2	0	0	3	1.0
23	F3b	1	0	0	0	1	0.3
24	F4b	0	2	0	0	2	0.7
25	M*	2	2	3	0	7	2.4
26	M12	3	2	0	0	5	1.7
27	M13	1	2	0	0	3	1.0
28	M17c	5	1	0	0	6	2.0
29	M20	6	2	3	0	11	3.7
30	M21a	1	3	0	1	5	1.7
31	M21c	2	1	0	0	3	1.0
32	M21d	1	1	0	0	2	0.7
33	M22a	0	2	0	0	2	0.7
34	M22b	0	0	0	1	1	0.3
35	M26a	3	0	1	0	4	1.3
36	M26b	0	1	1	0	2	0.7
37	M2b	0	1	0	0	1	0.3
38	M30	1	0	0	0	1	0.3
39	M32c	0	0	1	0	1	0.3
40	M37	0	0	0	1	1	0.3
41	M47	0	1	0	0	1	0.3
42	M5	0	1	0	0	1	0.3
43	M50	3	2	0	0	5	1.7
44	M51	3	0	1	0	4	1.3
45	M71	1	0	1	2	4	1.3
46	M72	1	0	0	1	2	0.7
47	M73	0	0	0	1	1	0.3
48	M74b	0	3	0	0	3	1.0
49	M77	1	0	0	0	1	0.3
50	M7b	1	2	0	7	10	3.4
51	M7b3	0	0	2	0	2	0.7
52	M7c3c	8	3	3	0	14	4.7

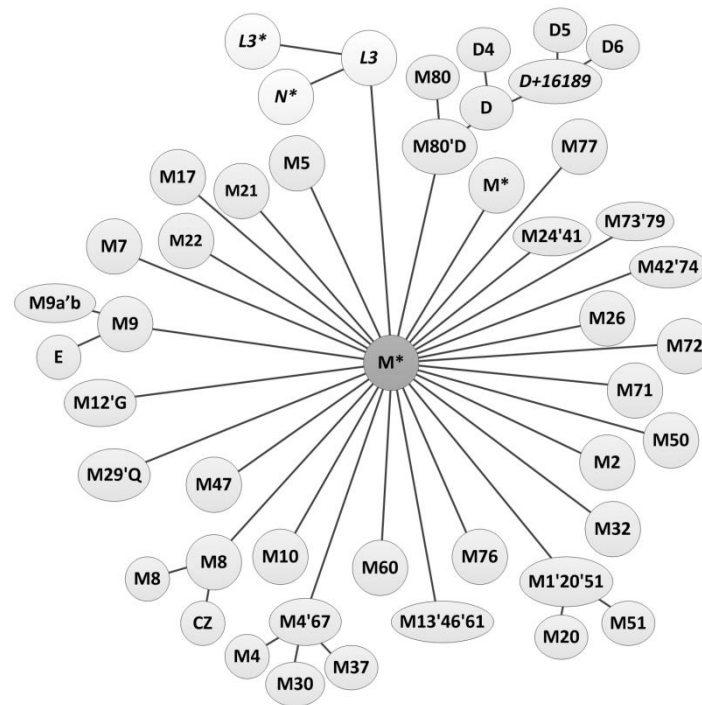
No.	Haplogroup	NEM	NWM	SEM	SWM	Total	%
53	N10	2	3	0	0	5	1.7
54	N21	2	1	0	1	4	1.3
55	N22	1	0	0	1	2	0.7
56	N8	0	0	1	0	1	0.3
57	N9a	1	0	0	1	2	0.7
58	P1d	1	0	1	0	2	0.7
59	Q1	0	0	1	0	1	0.3
60	Q3	0	1	0	0	1	0.3
61	R*	0	5	0	0	5	1.7
62	R11b	2	0	0	0	2	0.7
63	R21	1	0	0	0	1	0.3
64	R22	2	1	5	0	8	2.7
65	R6a	1	0	0	0	1	0.3
66	R7a	0	0	0	1	1	0.3
67	R9b	4	0	0	0	4	1.3
68	U1a	1	0	0	0	1	0.3
69	U2b1	0	1	0	0	1	0.3
70	U7	0	3	0	0	3	1.0
71	Y2a	0	5	1	0	6	2.0
	Total	109	98	56	34	297	100.0

The 297 ‘Malay ZZ’ samples, provided by Zafarina Zainuddin (248 of which were reported in Nur Haslindawaty *et al.*, 2010), were sequenced at the control region and compared with the HVS-I networks adapted from Hill (2005). In order to identify these Malay data from the existing ones in the networks, the data is denoted as “Malay ZZ” in the networks. Table 3.4 shows the HVS haplogroup distribution of Peninsular Malay according to sampling regions. As expected, there are 71 haplogroups including the M\* and R\* lineages in the Malay samples (now rather few compared with earlier MSEA studies, due to improved resolution and characterisation), more than three times the number of haplogroups among OA. The largest single haplogroup identified is B4c1b2 at 7.1%, followed by B5a and F1a1a, which both contributed 5.7% each to the pool. M7c3c is found in 4.7% of the Peninsular Malay, 3.7% each of M20 and B4a, and 3.4% each of E1a1a and M7b.

A total of 226 samples were selected for complete mtDNA genome sequencing: 19 Semang, 8 Senoi, 13 Aboriginal Malay and 186 Peninsular Malay (Appendix C – also with a column showing the world regional distributions of each complete sequence haplogroup). Representative samples from each haplogroup were chosen for complete mtDNA sequencing (See Section 2.1.1). The higher resolution phylogenetic trees of complete mtDNA genome variations in Peninsular Malaysia encompassed all three main Non-African haplogroups M, N, and R, apart from the East African M1. Haplogroup M including haplogroups M4’67, M5, M24, M47, M10, M60, M76, M7, M8’CZ, M9ab’E, M12’G, M13’46’61, M17, M2, M21,

M22, M26, M1'20'51, M32, M29'Q, M50, M71, M72, M42'74, M73'79, M77, M80'D, and novel M\*. Haplogroup N are N8, N10, N11, N21, N22, N9ab'Y and A. Haplogroup R including haplogroups R6, R7, R23, R30, R9b'cF, B4'5, R11'B6, R12'21, R22, R\*, and P. Haplogroup U is also present at lower levels in Peninsular Malaysia.

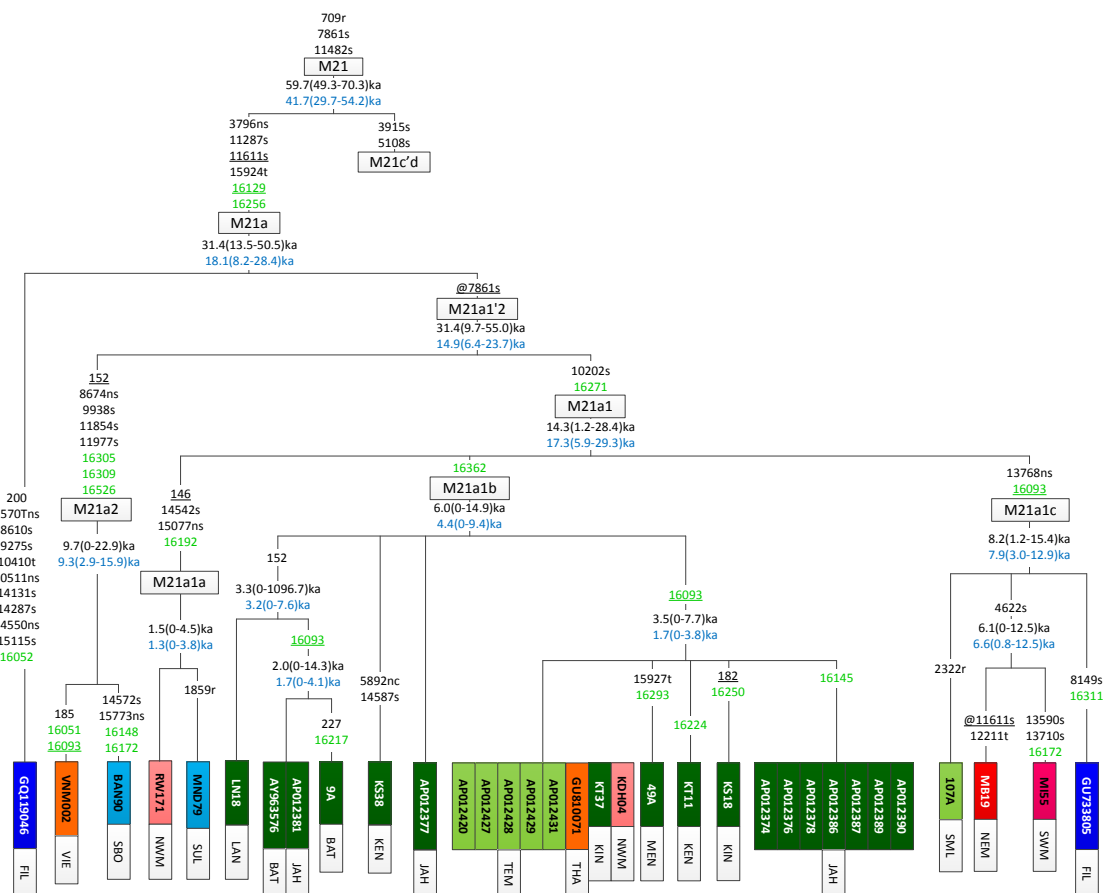
### 3.2 Haplogroup M



**Figure 3.7 Schematic diagram of haplogroup M's major subclades present in Southeast Asia.**

### 3.3 Haplogroup M21

Haplogroup M21 is among one of the first basal haplogroups found by Hill (2005). Figure 3.8 shows the phylogenetic tree of complete mtDNA sequences for haplogroup **M21a** a sub-group of **M21**, (the latter's age estimated at ~60 ka). Phylogeny of M21 includes 57 complete sequences, 33 M21a, eight M21c and 16 M21d. Here I propose new nomenclatures for the M21 subclades. M21 splits into two branches, M21a and M21c'd (M21b has been reassigned to M13b). The deepest split in **M21a**, dating to ~31 ka, is seen between the Philippines (one haplotype: Tabbada *et al.*, 2010) and largely Sunda populations, at ~ 31 ka. **M21a1'2** dates to ~31 ka and divides into M21a1 and M21a2 within the former Sunda continent.



**Figure 3.8** The tree of haplogroup M21a. Time estimates shown for clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. (BAT – Semang Batek, FIL – Philippines, JAH – Semang Jahai, KEN – Semang Kensiu, KIN – Semang Kintak, LAN – Semang Lanoh, MEN – Semang Mendriq, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SBO – South Borneo, SML – Aboriginal Malay Semelai, SUL – Sulawesi, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, THA – Thailand, VIE – Vietnam)

**M21a1** has a coalescence time estimated to ~14 ka, and is very frequent in Semang groups and also found in Senoi and Aboriginal Malays, although it is also found at low rates

throughout Southeast Asia. The largest sub-clade, **M21a1b** dates to ~6 ka and contains all the Semang, and Aboriginal Malay complete sequences M21a branches; and consists mostly of these (apart from two others, a Northwest Malay and a Thai representative). The young age of this ‘ethnically defined’ aboriginal cluster within a regionally ancient lineage (M21a), could imply recent drift and/or a local Holocene founding event (less likely) with ancestry found among all Aslian speakers. The first subclade nested within M21a1b, defined by a transition at np 152, dating to ~3 ka and is seen in the Semang Lanoh. Its subclade, defined by a recurrent mutation at np 16093, dating to ~2 ka, and is seen in the Semang Batek (Macaulay *et al.*, 2005 and this study) and Jahai (Jinam *et al.*, 2012). Another subclade nested within M21a1b, also defined by a recurrent mutation at np 16093, dating to ~4 ka. It is found mainly in the Semang Jahai (Jinam *et al.*, 2012), Kensiu, Kintak, Mendriq (this study), the Aboriginal Malay Temuan (Jinam *et al.*, 2012), Northwest Peninsular Malay (this study) and Thailand (Pradutkanchana, Ishida and Kimura, 2010).

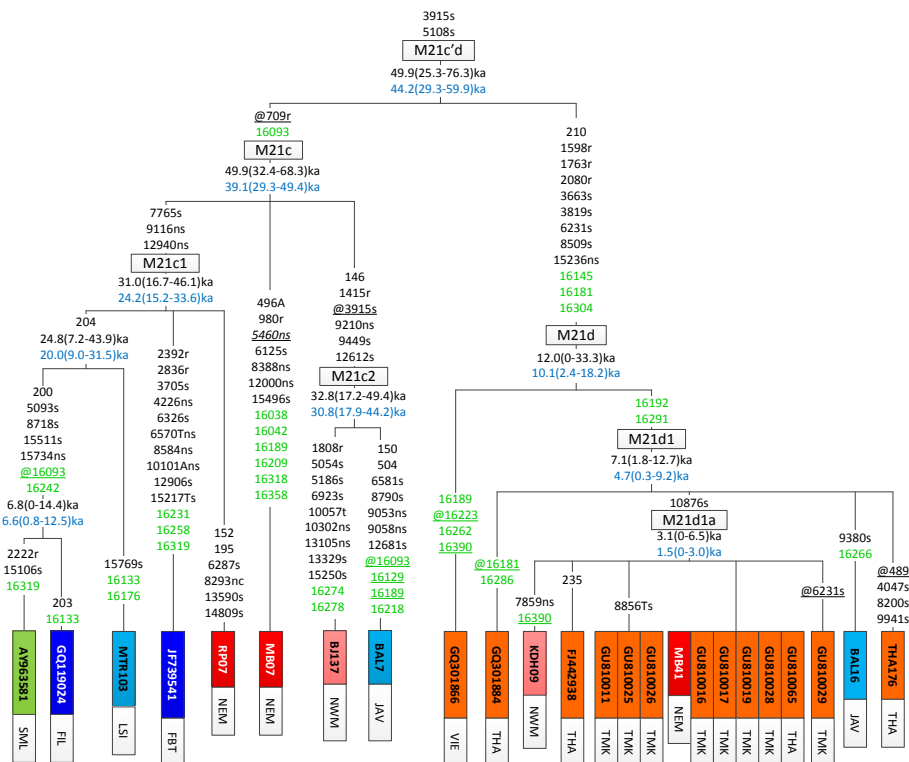
The two other branches of M21a1 have a much wider distribution in Southeast Asia, and only one representative among *Orang Asli*: **M21a1a** dates to ~2 ka, and is found in a Northwest Peninsular Malay and one individual in Sulawesi, Indonesia. **M21a1c**, dating to ~8 ka, and is seen in an Aboriginal Malay Semelai, two (Northeast and Southwest) Peninsular Malay, and a Filipino (Gunnarsdóttir *et al.*, 2011a). **M21a2** dates to ~10 ka, and is found in Vietnam and Banjarmasin of South Borneo, Indonesia (data from the Archaeogenetics Research Group, Huddersfield). It would appear more likely from this phylogeographic pattern that M21a1’2 (and M21a) formerly had a wider Sunda distribution, along with M21c and M21d, than that it first spread out from *Orang Asli* populations.

The whole-mtDNA M21a tree shows that the predominant Semang and Aboriginal Malay subclade (M21a1b) dates to ~6 ka. It remains highly localised within the *Orang Asli* mtDNAs in Peninsular Malaysia due to drift. However, the ancient ancestry appears to be captured by the relict descendant in other populations, such as the Peninsular Malay, Vietnam, Thailand, South Borneo, and Sulawesi of the Sunda shelf, and some recent offshoots to the Philippines.

**M21c’d** is a deep and widespread Sunda lineage, dating to ~50 ka, and can be divided into M21c and M21d. **M21c** is only detected once among *Orang Asli* from all the HVS-I data (Figure 3.10, shown also as a Holocene derived branch of M21c1 in the complete sequence Figure 3.9) and only in one Semelai, Aboriginal Malay (Hill, 2005). M21c has a coalescence



time of ~50 ka and is seen widely in MSEA, Indonesia and the Philippines, and in a Northeast Peninsular Malay. **M21c1** dates to ~31 ka, and is found in Northeast Peninsular Malay and the Philippines Batak (Scholes *et al.*, 2011). It then diverged ~25 ka and seen in Mataram of Lesser Sunda Islands, which subsequently a subclade nested within, dating to ~7 ka, seen in the Aboriginal Malay Semelai (mentioned above as HVS, complete sequence given in Macaulay *et al.*, 2005) and the Philippines (Tabbada *et al.*, 2011). The Philippines Batak is the so-called negrito group who predominantly lead a hunter-gatherer existence in small scattered settlements on the island of Palawan. The Batak speak languages of the recently spread Austronesian family, and they are believed to have replaced non-Austronesian languages spoken by negrito before the Holocene (Reid, 1994). The recent admixture of the Batak populations with neighbouring non-negrito Tagbanua tribe (Eder, 1987; Migliano *et al.*, 2007) has accelerated the disappearance of Batak physical and cultural distinctiveness. Meanwhile, the introduction of agriculture into Batak hunting territory has effectively influenced their lifestyle (Eder, 1987).



**Figure 3.9** The tree of haplogroup M21c'd. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (FBT – Philippines Batak, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SML – Aboriginal Malay Semelai, THA – Thailand, VIE – Vietnam, TMK – Thailand Moken)

**M21d** dates to ~12 ka and is mainly found in MSEA – Vietnam and the Moken Sea Gypsies of Thailand. The Moken inhabit the Mergui Archipelago off the coast of Myanmar and Thailand by boats and subsist through maritime foraging (White, 1922; Sopher, 1965; Ivanoff, 2005). Their own language belongs to the Malayo-Polynesian branch of the Austronesian language family (Larish, 1999; Gordon, 2005). **M21d1** dates to ~7 ka, and the basal lineages are seen in Thailand and Java, Indonesia (Peng *et al.*, 2010; Archaeogenetics Research Group, Huddersfield). The Northern Peninsular Malay (this study) and Thai (Dancause *et al.*, 2009) formed a cluster with the majority of Moken sequences (Pradutkanchana, Ishida and Kimura, 2010) dating to ~3 ka.

Figure 3.10 shows the HVS-I network for haplogroup M21, which is very poorly resolved in comparison with the whole-mtDNA tree, as there are few informative HVS-I sites within the tree. In previous HVS-I studies, M21a is predominantly found in the Semang, also present in the Aboriginal Malays, Malay (Zainuddin and Goodwin, 2004; and “Malay ZZ”), Banjarmasin of South Borneo and the “Maniq” Semang of Southern Thailand (Fucharoen *et al.*, 2001), corresponding with M21a1b clade in the whole-mtDNA Figure 3.8. M21c, as mentioned, is only found in the Aboriginal Malays. I will describe M21b (currently M13b) later.

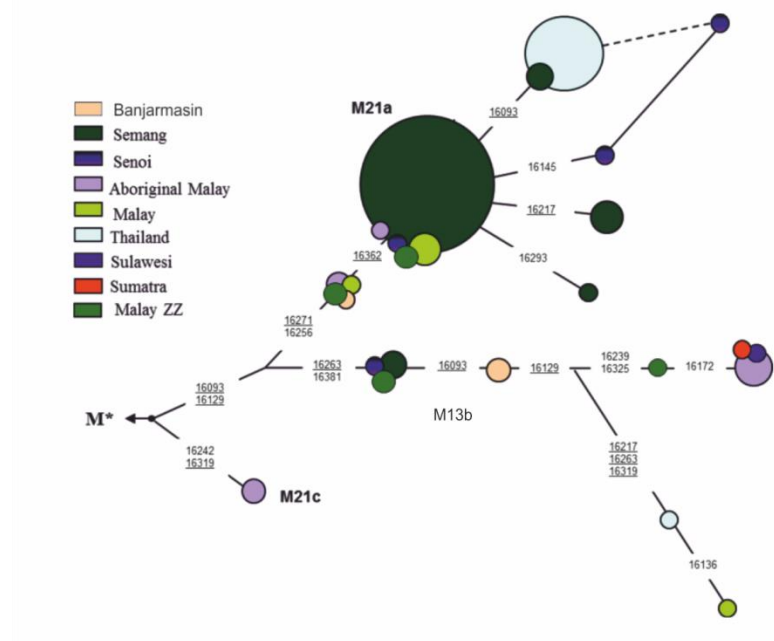


Figure 3.10 HVS-I network of M21\*. M21b has been reassigned to M13b. Figure adapted from Hill (2005).

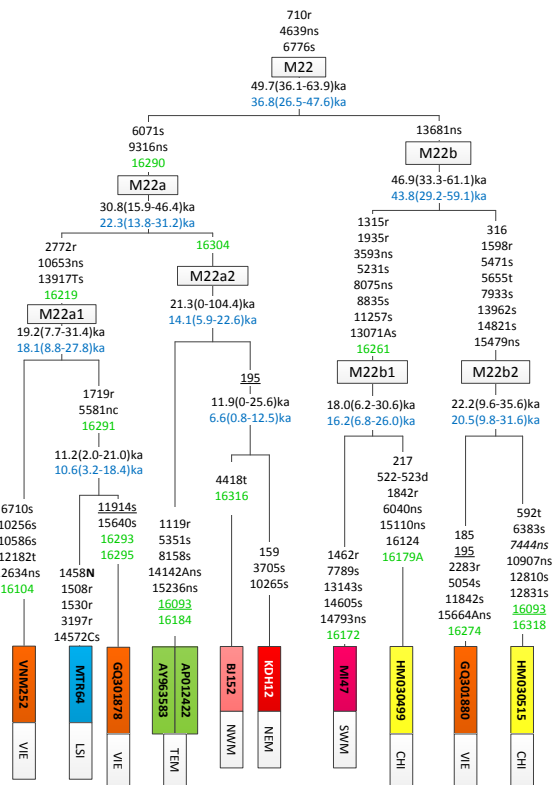
Our high-resolution tree and broader sampling has shown that M21 is a deep and ancient Sunda haplogroup, which do not just restricted to Peninsular Malaysia and South Borneo. The relict descendants have a widespread distribution on the Sunda shelf and as east as the Philippines (which was not part of Sundaland), which almost certainly suggest a Pleistocene Sunda origin.

### 3.4 Haplogroup M22

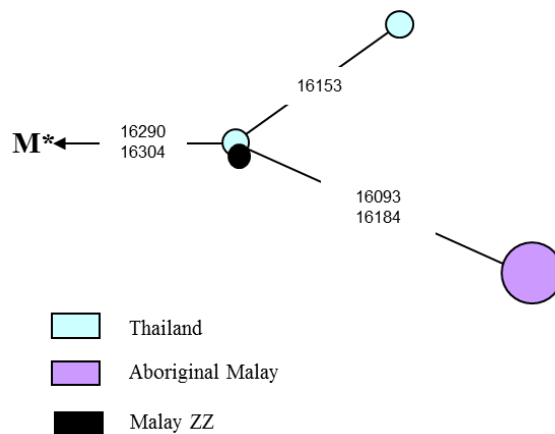
M22 is basal M haplogroup that is previously found predominantly in Temuan Aboriginal Malays and Thai (Hill *et al.*, 2006). M22 dates to ~50 ka and it is diverged into M22a and M22b, and is restricted mainly to relict regions of the former Sunda continent with two derived haplotypes in South China. The phylogeny of M22 includes eleven complete sequences: seven M22a and four M22b, with new nomenclature for its subclades (i.e. M22a1, M22a2, M22b1 and M22b2). M22 appears, from the phylogeny, to be an indigenous haplogroup from MSEA, where it is seen in China, Vietnam, the Aboriginal Malays, Peninsular Malay, and Java, Indonesia.

**M22a** again divides into M22a1 and M22a2 ~31 ka (Figure 3.11). **M22a1** dates to ~19 ka in Vietnam (Archaeogenetics Research Group, Huddersfield), and subsequently at ~11 ka found in Vietnam and Mataram of Lesser Sunda Islands, Indonesia (Peng *et al.*, 2010). **M22a2**, dating to ~21 ka, is seen in the Temuan (Macaulay *et al.*, 2005; Jinam *et al.*, 2012), with a subsequent branch found among the Northern Peninsular Malay dating to around 12 ka.

**M22b** dates to ~47 ka and divides into M22b1 and M22b2 (Figure 3.11). **M22b1**, dating to ~18 ka, is seen in Southwest Peninsular Malaysia (this study), and South China (Kong *et al.*, 2011). **M22b2**, dating to ~22 ka, and is found in Guangdong, South China (Kong *et al.*, 2011) and Vietnam (Peng *et al.*, 2010). The whole-mtDNA tree seems to suggest an origin in MSEA and spread southwards into Java, Indonesia via Peninsular Malaysia.



**Figure 3.11** The tree of haplogroup M22. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, VIE – Vietnam)



**Figure 3.12** HVS-I network of M22. Figure adapted from Hill (2005).

Figure 3.12 below shows the HVS-I network of M22 showing that it was previously found in Thailand, the Aboriginal Malays (Hill *et al.*, 2006), and “Malay ZZ” (this study). The Thai individual with a transition at np 16153 would be nested within M22a2 in the whole-mtDNA tree. Hill *et al.* (2006) suggested that the root of M22 could be somewhere in MSEA, which is consistent with the whole-mtDNA tree.

### 3.5 Haplogroup M7

Haplogroup M7 is one of the most common Asian haplogroups in China, Korea, Japan, and Island Southeast Asia. M7 dates to ~56 ka, and divides into M7a and M7b-g, the latter is further divided into M7c'e'f and M7b'd'g (Figure 3.13). M7a is a Northeast Asian haplogroup, reported only in Japan and Korea (Kivisild *et al.*, 2002; Tanaka *et al.*, 2004). M7b-g dates to ~53 ka and is divided into M7g and M7b'd (including M7b and M7d), where the basal lineages are seen mostly in China. M7b and M7c have wider distribution in Southeast Asia as indicated in the figure below. M7 phylogeny includes 173 complete sequences: 47 M7a, 59 M7b, 60 M7c, one M7d, three M7e and two M7g.

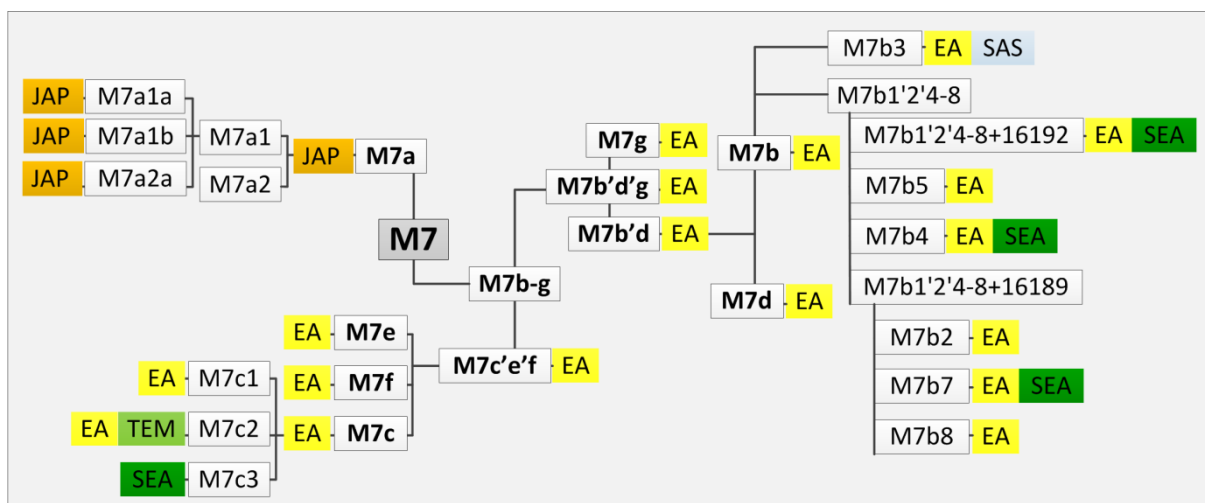


Figure 3.13 Schematic diagram of haplogroup M7 and its major subclades distribution. (EA – East Asia, JAP – Japan, SAS – South Asia, SEA – Southeast Asia, TEM – Aboriginal Malay Temuan)

#### 3.5.1 Haplogroup M7a

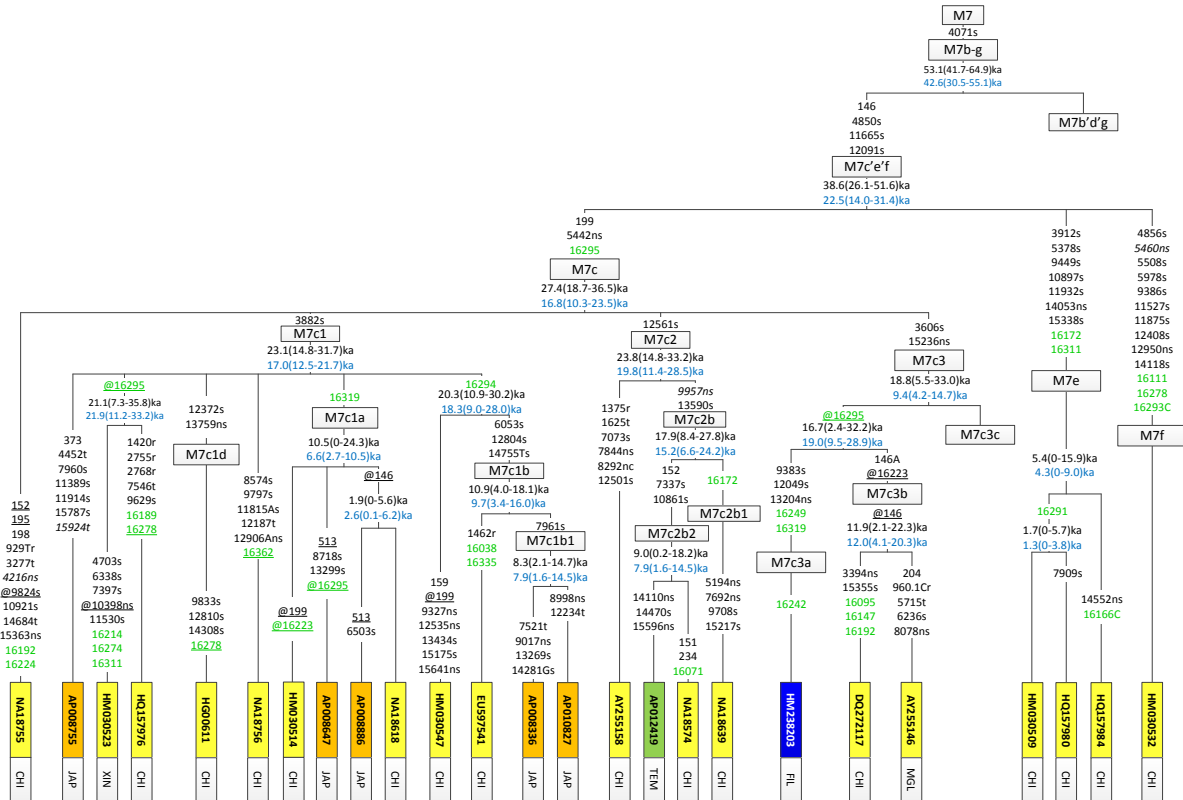
**M7a** is defined by variants at nps 2626, 2772, 4386, 4958, 12771 and 16209 with a divergence time of ~27 ka. M7a is entirely restricted to Japan, sampled from four locations in Japan (Tokyo, Chiba, Aichi and Gifu) reported by Tanaka *et al.* (2004) and Nohira *et al.* (2010). See Appendix E for more description on M7a.

#### 3.5.2 Haplogroup M7c'e'f

**M7c'e'f**, dating to ~39 ka, and is divided into M7c, M7e, and M7f (Figure 3.14). **M7c** dates to ~27 ka and its subclades are mostly seen in China (Yao *et al.*, 2002a; Yao *et al.*, 2002b; Hartmann *et al.*, 2009; Kong *et al.*, 2011; Peng *et al.*, 2011b; Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004; Nohira *et al.*, 2010), with the exception of subclade M7c3c that is widely distributed throughout SEA, and a single instance of Temuan Aboriginal Malay

nested within M7c2b2. **M7e** (dates to ~5 ka) and **M7f** is represented here by a single Chinese lineage, both are found in south China (Kong *et al.*, 2011; Peng *et al.* 2011b).

Haplogroup **M7c1** and two out of the three subclades are similarly dated to the LGM (Figure 3.14), mainly restricted to China (Hartmann *et al.*, 2009; Kong *et al.*, 2011) and Japan (Tanaka *et al.*, 2004; Nohira *et al.*, 2010).

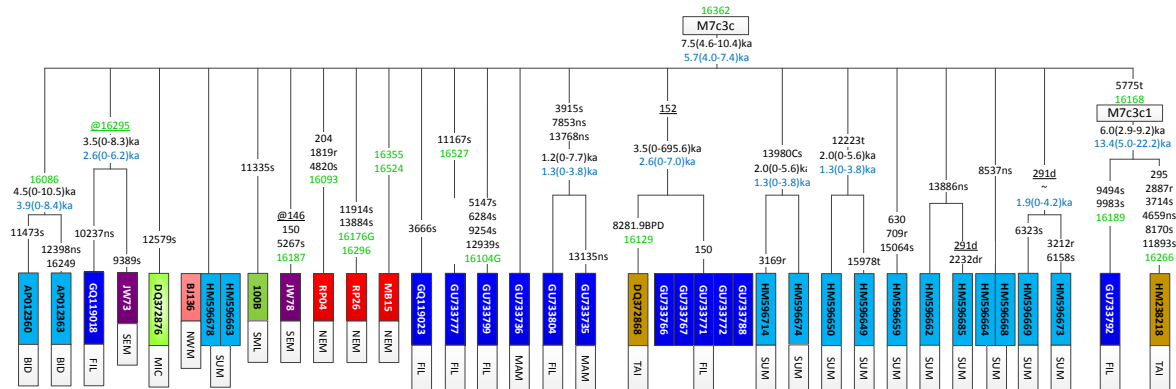


**Figure 3.14** The tree of haplogroup M7c'e'f excluding M7c3c. Time estimates shown for clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan, MGL – Inner Mongolia, China, TEM – Aboriginal Malay Temuan, XIN – Xinjiang, China)

**M7c2**, dating to ~24 ka, has a basal lineage seen in Liaoning, northeast of China (Kong *et al.*, 2003a). Its subclades are seen in north China (Zheng *et al.*, 2010), and nested within subclade M7c2b2 (dates to ~9 ka) is the Aboriginal Malay Temuan (Jinam *et al.*, 2012), clearly suggesting a northern origin and witness to a small-scale late glacial dispersal south into the Malay Peninsula.

**M7c3** dates towards the end of the LGM at ~19 ka and is one of the most commonly found haplogroups in ISEA (Hill *et al.*, 2007). Minor subclades **M7c3a** and **M7c3b** are nested within a subclade defined by a back mutation at np 16295, and dates to ~17 ka. M7c3a is so far only represented by a single instance from the Philippines (Loo *et al.*, 2011), and

M7c3b, dating to ~12 ka, appears to be confined in Guizhou, South China and Inner Mongolia (Kong *et al.*, 2003a; Kong *et al.*, 2006).



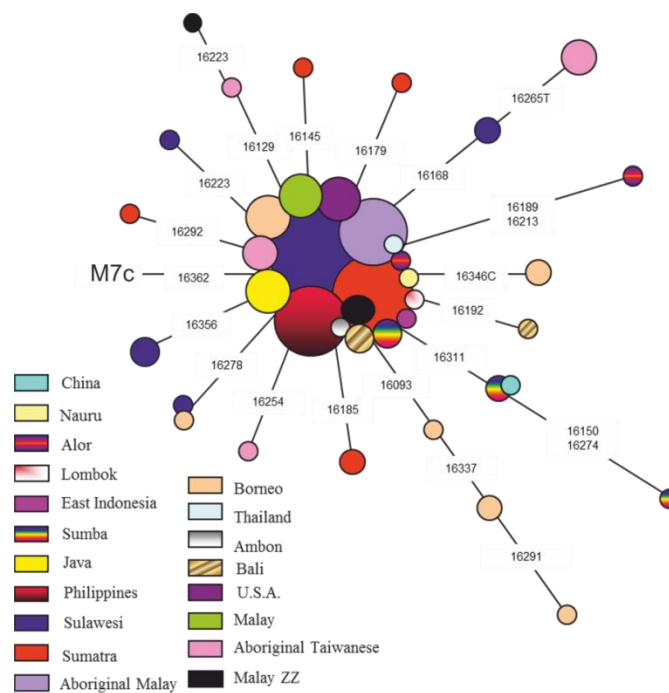
**Figure 3.15** The tree of haplogroup M7c3c. Time estimates shown for clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. Mutation at np 310 in the GU733- sequences (Gunnarsdóttir *et al.*, 2011a) was removed from the tree for posing incorrect evolutionary pathways. (BID – Bidayuh Sarawak, FIL – Philippines, MIC – Micronesia, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUM – Sumatra, TAI – Taiwan)

**M7c3c** (previously called M7c1c) dates to ~7.5 ka (Figure 3.15) and is widely distributed throughout ISEA in Peninsular Malaysia, Sumatra, Borneo and the Philippines, and elsewhere in the Aboriginal Taiwanese, and east in Majuro Atoll, Micronesia (Pierson *et al.*, 2006; Tabbada *et al.*, 2010; Gunnarsdóttir *et al.*, 2011a; Gunnarsdóttir *et al.*, 2011b; Loo *et al.*, 2011; Jinam *et al.*, 2012). In Malaysia, it is seen in the Aboriginal Malay (but Aslian-speaking) Semelai and modern Malay in the northern and southeastern Peninsular Malay (this study). A subclade of M7c3c defined by a transition at np 16086 (~5 ka) has recently found among the Sarawak Bidayuh of north Borneo (Jinam *et al.*, 2012). A subclade defined by a reversion at np 16295, dating to ~4 ka, is seen in Filipino (Tabbada *et al.*, 2010) and a Southeastern Peninsular Malay. One negrito Mamanwa from Philippines shares a subclade with an urban Filipino (Gunnarsdóttir *et al.*, 2011a) that diverged at ~1 ka. A possible subcluster characterised by a transition at the fast site np 152 dates to ~4 ka, is shared between one Taiwanese aboriginal (Pierson *et al.*, 2006) and the Monobo and Surigaonons of Philippines (Gunnarsdóttir *et al.*, 2011a).

Additionally, there are at least four subclades with an average date of ~2 ka formed by the Sumatran of Indonesia (Gunnarsdóttir *et al.*, 2011b). The first two subclusters have a similar estimated age of ~2 ka. The third and fourth subclusters are not dated because they are defined by deletions that are not used in the age estimations. **M7c3c1** dates around 6 ka,

and is found in Aboriginal Taiwanese (Loo *et al.*, 2011) and the Philippines (Gunnarsdóttir *et al.*, 2011a).

Similar to M7b, therefore, M7c3c offers a possible origin in the postulated Austronesian-speaking dispersal from South China and Taiwan through ISEA into Peninsular Malaysia (Bellwood, 1997), although the point estimates for the ages of M7c3c and M7c3c1 seem too old for that archaeo-linguistic model, and the phylogeography of the M7c3c1 clade in the whole-mtDNA tree could also be consistent with a Philippines/ISEA origin and a reverse migration to Taiwan quite recently.



**Figure 3.16 HVS-I network of M7c3c. Figure adapted from Hill (2005).**

The HVS-I data in Figure 3.16 confirm that M7c3c is common throughout Indonesia, Peninsular Malaysia, Thailand, Borneo, the Philippines and the Aboriginal Taiwanese. The derivatives defined by a transition at np 16168 are seen in Sulawesi, which is recognisable as M7c3c1 on the whole-mtDNA tree, and a further transversion at np 16265T found in East Indonesia would have nested within the same subclade, which suggests M7c3c1 is not only restricted to Taiwan and the Philippines (whole-mtDNA tree), but also present in Sulawesi and East Indonesia.

The whole-mtDNA tree and HVS-I data of M7c appears to indicate an ultimate origin in China. However, subclade M7c3c predates the traditional Neolithic Out of Taiwan model to mid-Holocene period ~7.5 ka. The genetic diversity in Taiwan and Borneo appears to be



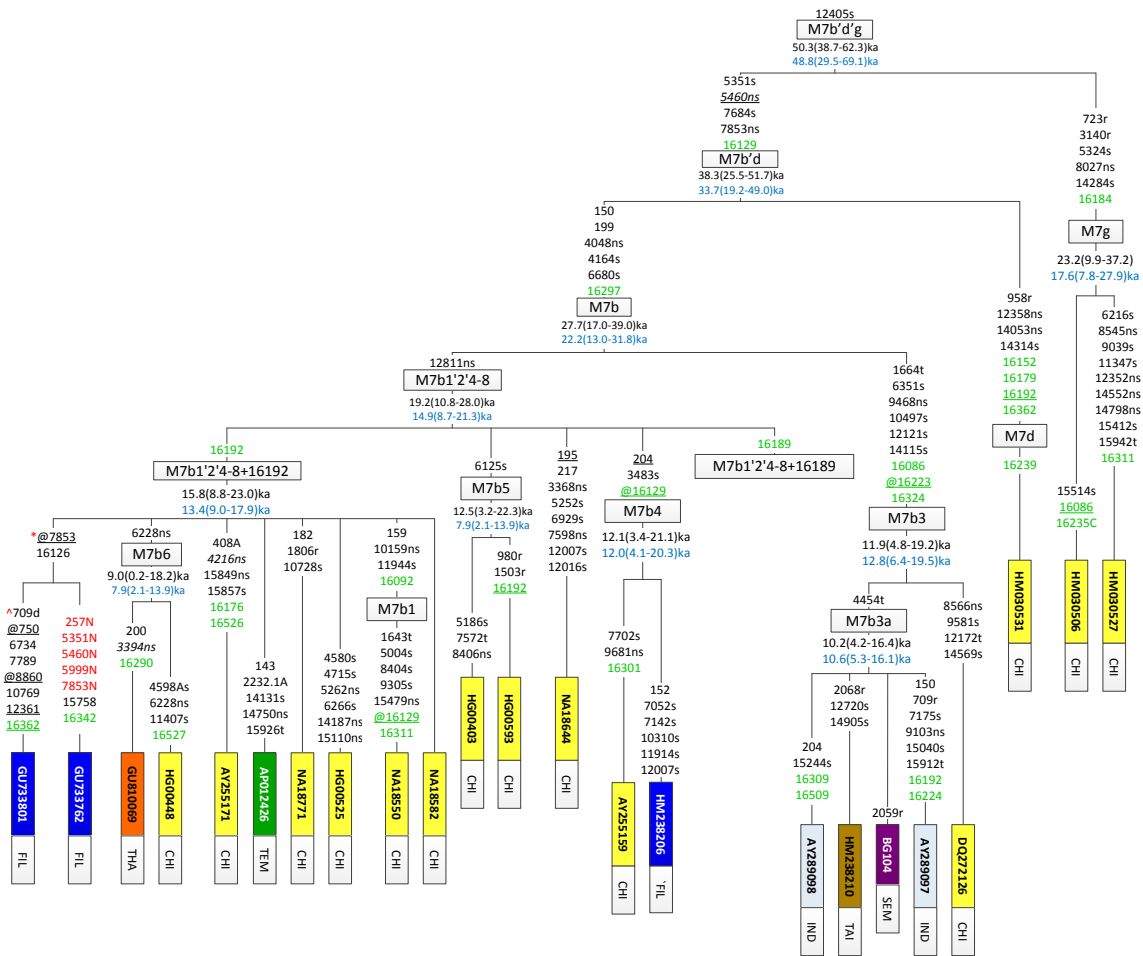
more diverse than would be expected if it had arrived from China <6 ka (Hill *et al.*, 2007). The spatial frequency distribution for haplogroup M7c3c (M7c1c in Hill *et al.*, 2007) indicates that it is more centered on Borneo and Sulawesi. The new analysis confirmed the finding in Hill *et al.* (2007) that although M7c3c is possibly a strong marker for Out of Taiwan dispersal, the older date obtained from the whole-mtDNA analysis shows it has a postglacial mid-Holocene dispersal in ISEA, which is likely to centre on Borneo and reverse migration to Taiwan and travel as far east as Micronesia.

### 3.5.3 Haplogroup M7b'd'g

**M7b'd'g** dates to ~50 ka. It is divided into M7b'd and M7g, dating to ~38 ka and ~23 respectively. M7d and M7g are rarer haplogroups seen in China. **M7d** is represented here by a single instance in Qinghai, China, and **M7g**, which dates to ~23 ka, are found in Sichuan and Guizhou of China (Kong *et al.*, 2011).

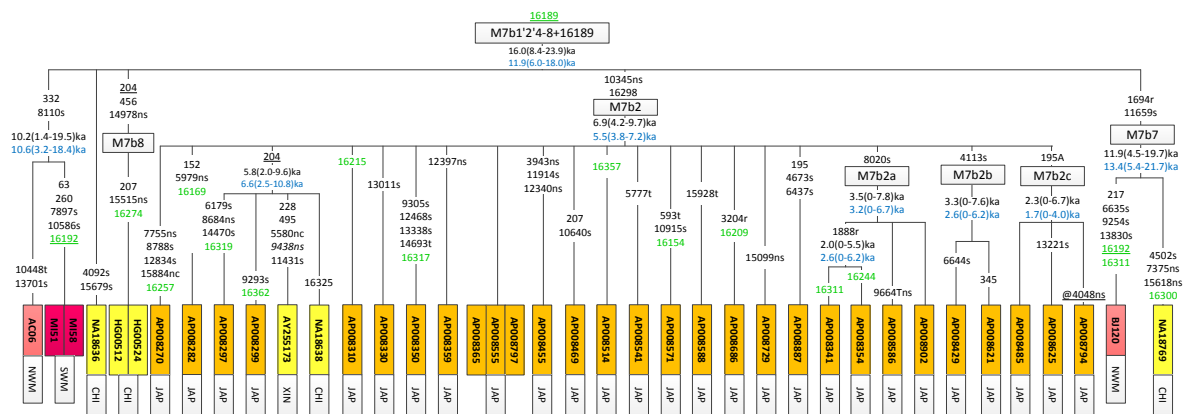
**M7b**, dating to ~28 ka, and is divided into haplogroups M7b1'2'4-8 and M7b3, where M7b1'2'4-8 encompasses four subclades as shown in Figure 3.17. **M7b1'2'4-8+C16192T** dates to ~16 ka and the basal lineages are mostly seen in northern and southern China (Kong *et al.*, 2003a; Zheng *et al.*, 2011), and a lineage from the Temuan Aboriginal Malay (Jinam *et al.*, 2012). **M7b6**, dating to ~9 ka, is seen in south China (Zheng *et al.*, 2011) and Thailand (Pradutkanchana, Ishida and Kimura, 2010). A subclade formed by the Philippines samples (Gunnarsdóttir *et al.*, 2011a) is shown in the tree, but they were excluded from age estimations because of the ambiguous sites and gaps present in these sequences. **M7b5** dates to ~13 and seen in south China (Zheng *et al.*, 2011). **M7b4** dates to ~12 ka, and is found in Hunan China (Kong *et al.*, 2003a) and the Ivatan Philippines (Loo *et al.*, 2011). Ivatan Islanders are Austronesian speakers from Orchid Island and the Batanes archipelago located between Taiwan and the Philippines, who have a close cultural relationship with the Yami tribe in Taiwan, which is the only non-Formosan Austronesian speakers among Taiwan Aborigines (Blust, 1999). However, it is not possible to infer with just a few complete sequences.

**M7b3** dates to ~12 ka and a basal lineage is seen in Guizhou, China (Kong *et al.*, 2006), nested within a subclade M7b3a that dates to mid-Holocene ~10 ka and seen in Southeast Peninsular Malaysia, Yami of Aboriginal Taiwanese and India (Ingman and Gyllensten, 2003).

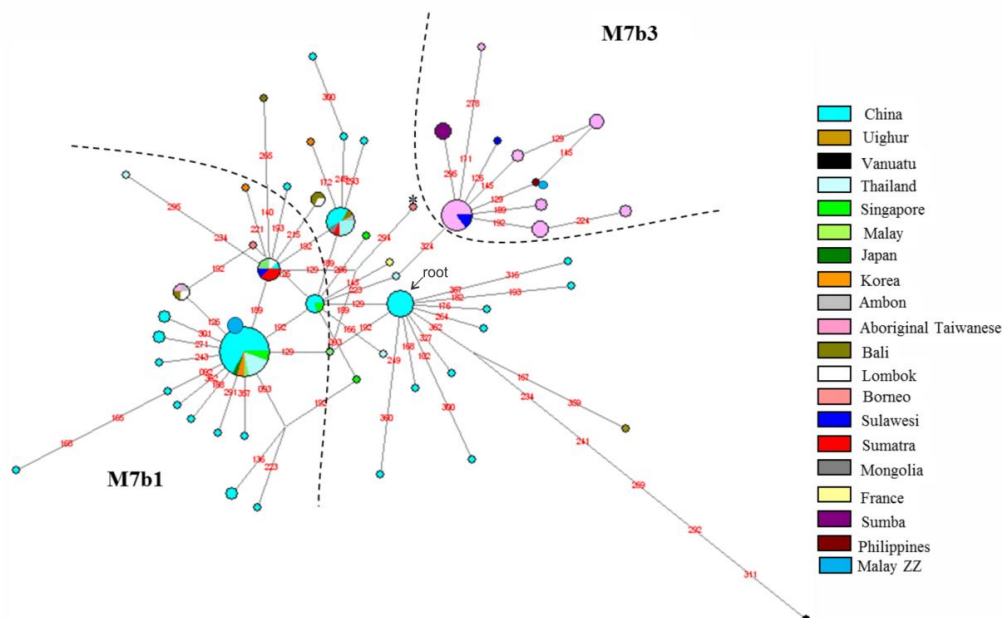


**Figure 3.17** The tree of haplogroup M7b'd'g (excluding M7b1'2'4-8+16189). Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. The Philippines sequences marked by “\*” are excluded from age estimations. (CHI – China, FIL – Philippines, IND – India, SEM – Southeast Peninsular Malay, TAI – Taiwan, TEM – Aboriginal Malay Temuan, THA – Thailand)

A subclade of **M7b1'2'4-8** defined by a transition at 16189 dates to ~16 ka and can be divided into M7b2, M7b7, and M7b8 (Figure 3.18). M7b2, dating to ~7 ka, is mostly seen in Japan (Tanaka *et al.*, 2004) with some in China (Kong *et al.*, 2003a; Zheng *et al.*, 2011). There are two other rare subclades, **M7b7** dates to ~12 ka and is found in one Northwest Peninsular Malay (this study) and one in Beijing China (Zheng *et al.*, 2011), and **M7b8** is represented by two similar instances seen in South China (Zheng *et al.*, 2011). Also nested within subclade M7b1'2'4-8 is a branch defined by transitions at nps 332, 8110 and 16189, dating to ~10 ka, and found in Northwest and Southwest Peninsular Malay only (this study).



**Figure 3.18** The tree of haplogroup M7b1'2'4-8. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NWM – Northwest Peninsular Malay, SWM – Southwest Peninsular Malay, XIN – Xinjiang, China)



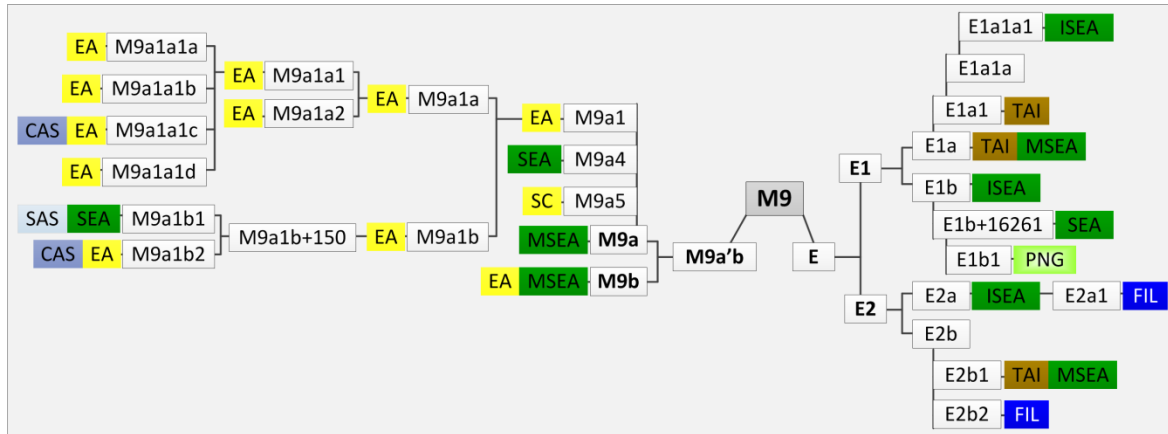
**Figure 3.19** Weighted HVS-I network of M7b\*, M7b1 and M7b3 types with the root indicated. Figure adapted from Hill (2005).

Figure 3.19 shows the HVS-I network of M7b (subclades M7b1 and M7b3 are recognisable and outlined in the network). Although at much lower resolution and lesser informative sites, the HVS-I data corresponds the whole-mtDNA tree where it shows M7b is most commonly reported in South China by Kivisild *et al.* (2002), Yao *et al.* (2002a) and Yao *et al.* (2002b). Elsewhere, M7b is seen at decreasing levels in Singapore, the Philippines, Thailand and Vanuatu (Hill, 2005). The HVS-I of M7b3 haplotypes are mainly found in Taiwanese Aborigines, and less common in Toraja Sumatra, the Philippines and Peninsular Malay, much similar to what is observed in the new analysis.

The whole-mtDNA result suggests M7b has a pre-LGM origin in China seen, in particular, from the diversity of the branching order in China, with Late Glacial expansions on the Sunda shelf as evident by the extant relict lineages in SEA. The M7b Temuan and Malay lineages (of different subclades) appear to have maternal origin in South China during the late Pleistocene (~16 ka) and early-Holocene dispersals (~12-10 ka) into Peninsular Malaysia, instead of a Neolithic event coming from offshore.

### 3.6 Haplogroup M9

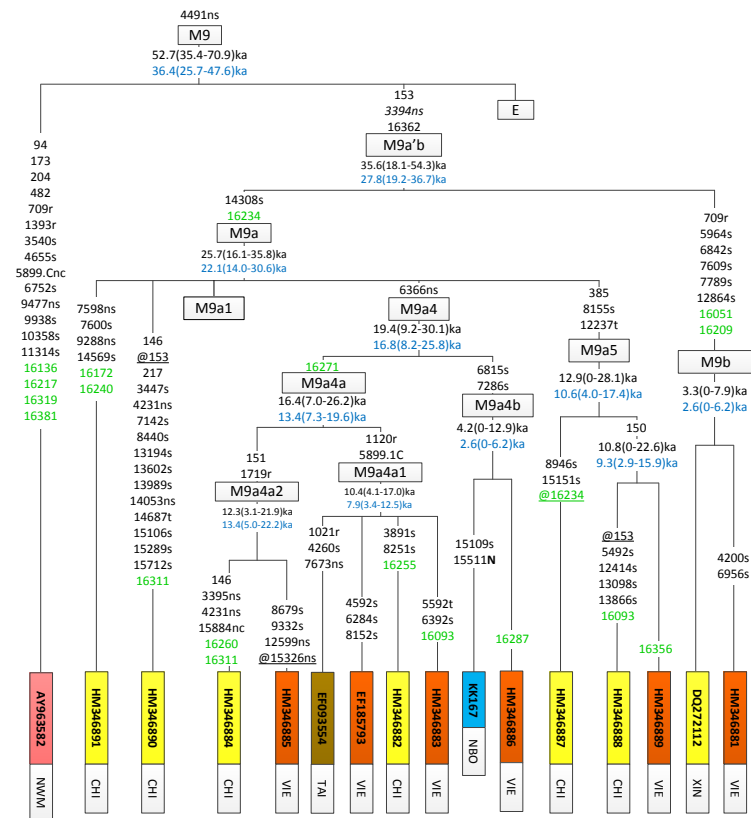
Haplogroup **M9**, dating to ~53 ka, includes haplogroup M9a'b and E. E dates to ~29 ka and M9a'b to ~36 ka. The tree of M9 includes 200 complete sequences: 79 M9a'b and 121 E. The complete sequence tree of M9a'b suggests a root for M9a'b in South China (Figure 3.20), however, there is a basal paraphyletic branch of M9 from the Northeastern Malay Peninsula, suggesting an older root there for M9 with spread northwards into South China and elsewhere in East Asia (see also discussion in Soares *et al.*, 2008). M9a has basal lineages in South China, which diversify in China and Japan as M9a1 and M9a5, and in South China and SEA as M9a4. M9a1a and its subclades are absent in SEA, but are commonly seen in China and Japan, except that M9a1a1c is also seen in Central Asia (see discussion in Peng *et al.*, 2011a). M9a1b suggests a root in China and spread to Japan, while its subclade M9a1b1 is seen in South Asia and SEA, and M9a1b2 is found in Central and East Asia (see discussion in Peng *et al.* 2011a). Haplogroup E is a primary branch of M9 and is divided into E1 and E2, each subdivided into 2 subclades, E1a, E1b, E2a and E2b. All 4 subclades have a very distinctive geographic distribution in ISEA, which is highly informative about the demographic history of the region. However, only E1a and E2b are found in Taiwan. E1b and E2a are both largely confined to ISEA, although occasionally extending to New Guinea and Peninsular Malaysia, indicating that both arose in ISEA and dispersed fairly recently east and west.



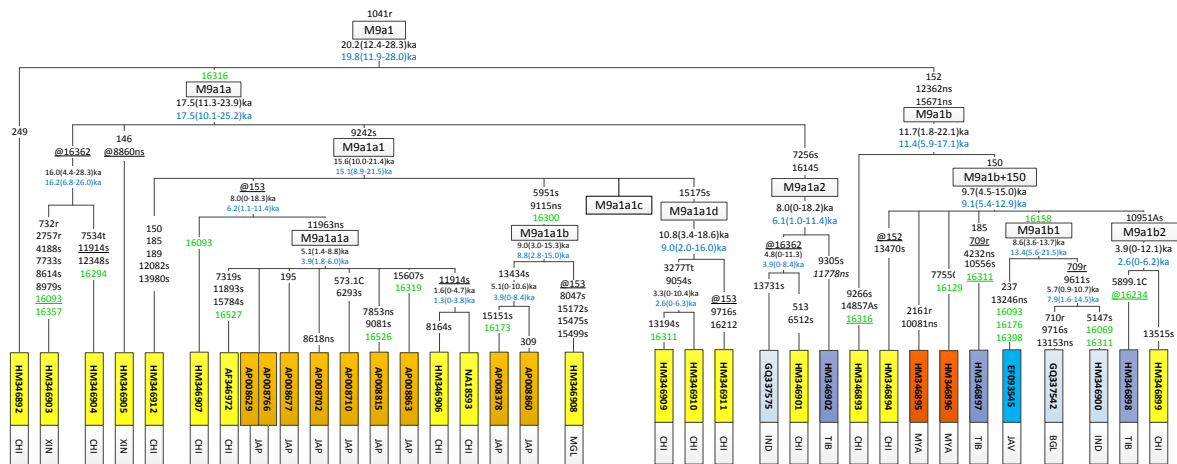
**Figure 3.20 Schematic diagram of haplogroup M9 and its major subclades distribution.** (CAS – Central Asia, EA – East Asia, FIL – Philippines, ISEA – Island Southeast Asia, MSEA – Mainland Southeast Asia, PNG – Papua New Guinea, SAS – South Asia, SC – South China, SEA – Southeast Asia, TAI - Taiwan)

### 3.6.1 Haplogroup M9a'b

Figure 3.21 shows that there is a single paraphyletic basal lineage within **M9**, found in a Northwest Peninsular Malay (Macaulay *et al.*, 2005) – although its assignment to M9 relies on a single variant at np 4491 (Soares *et al.*, 2008), a site which changes six times in the global phylogeny of Soares *et al.* (2009). **M9a'b** dates to ~36 ka, and bifurcates immediately into the major M9a and the much rarer M9b, which both have representatives in both China and the Sunda continent (mainly Vietnam, in M9a4). **M9a** dates to ~26 ka, and consists of subclades M9a1, M9a4 and M9a5, which are dated to ~20 ka, ~19 ka and ~13 ka respectively. Given the basal MSEA lineage within M9 and the majority presence of MSEA representatives in the primary M9a4 branch, the precise geographic origins of M9a'b, M9a and M9b (i.e., whether South China or MSEA) are open to question (as discussed in Peng *et al.*, 2011a; see also Soares *et al.*, 2008). However, basal types of M9a are found in South China and Southeast Asian lineages are largely restricted to Vietnam and Myanmar, suggesting Holocene gene flow from a South China source into the latter parts of MSEA and sporadically also Taiwan/Indonesia. The database of HVS variation (which identifies both major branches of M9a'b) confirms that the centre of gravity for the distribution is strongly South China, with few examples further to the south, although there is a very heavy skew towards the M9a1a subclade, which is not seen south of China.



**Figure 3.21** The tree of haplogroup M9a'b, excluding M9a1. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NBO – North Borneo, NWM – Northwest Peninsular Malay, TAI – Taiwan, VIE – Vietnam, XIN - Xinjiang)



**Figure 3.22** The tree of haplogroup M9a1, excluding M9a1a1c. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BGL – Bangladesh, CHI – China, IND – India, JAP – Japan, JAV – Java, Indonesia, MGL – Inner Mongolia, China, MYA – Myanmar, TIB – Tibet, XIN – Xinjiang)

**M9a4**, dating to ~16 ka, is rare and divided into M9a4a and M9a4b. **M9a4a** includes two subclades: **M9a4a1** and the newly named **M9a4a2**. M9a4a1 dates to ~10 ka and is found in individuals in Taiwan, Vietnam and South China (Soares *et al.*, 2008; Peng *et al.*, 2011a).

M9a4a2, dating to ~12 ka, is seen in South China and Vietnam (Peng *et al.*, 2011a). M9a4b dates to ~4 ka and seen in Vietnam (Peng *et al.*, 2011a) and North Borneo (Archaeogenetics Research Group, Huddersfield). **M9a5** is seen in Southern China, and a nested subclade dates to ~11 ka, and is found in China and Vietnam (Peng *et al.*, 2011a). **M9b** dates to ~3 ka and is represented in the tree by only two individuals, from Xinjiang China (Kong *et al.*, 2006) and Vietnam (Peng *et al.*, 2011a).

In Figure 3.22, **M9a1** dates to ~ 20 ka and has a basal lineage found in South China, which can be divided into M9a1a and M9a1b, both showing post-glacial expansion centred largely in China. **M9a1a** is by far the most frequent subclade within M9a1b, and also appears to be an East Asian haplogroup, seeing as it is mainly found in South China, Japan, Korea, and at relatively lower frequency in Tibet, India, and Mongolia (Ingman *et al.*, 2000; Kong *et al.*, 2003a; Tanaka *et al.*, 2004; Kong *et al.*, 2006; Soares *et al.*, 2008; Peng *et al.*, 2011a; Zheng *et al.*, 2011). **M9a1a** dates to ~17.5 ka, and has at least three subclades, M9a1a+@16362, M9a1a1, and M9a1a2. **M9a1a+@16362**, a loss of transition at np 16362 from M9a1a, has been dated to ~16 ka, and detected in single individuals from Xinjiang and South China (Peng *et al.*, 2011a). **M9a1a1** dates to ~16 ka, and is subsequently divided into M9a1a1a, M9a1a1b, M9a1a1c and M9a1a1d. A reversion at np 153 generates a putative pre-M9a1a1a node, dating to ~8 ka in northern China. **M9a1a1a** dates to ~5 ka and is confined to China and Japan only (Ingman *et al.*, 2000; Tanaka *et al.*, 2004; Peng *et al.*, 2011a; Zheng *et al.*, 2011). An additional transition at np 11914 defines a subclade of M9a1a1a dating to ~2 ka and detected in northern China (Peng *et al.*, 2011a; Zheng *et al.*, 2011). M9a1a1c is shown in Appendix E.

**M9a1a1b** dates to ~9 ka. It is shared between an individual from Inner Mongolia in China (Peng *et al.*, 2011a) and two Japanese samples (Tanaka *et al.*, 2004). The latter belong to a subclade dating to ~5 ka, similar to the Japanese subclade M9a1a1a, pointing to a colonisation event ~5 ka involving both subclades as founders. **M9a1a1d** dates to ~ 11 ka, and is found only in three individuals from northwest and southwest China (Peng *et al.*, 2011a).

**M9a1b** has a basal lineage in China, with a nested subclade M9a1b+150 (~10 ka) is seen across in China, Myanmar, Tibet, India and Bangladesh, suggesting a Holocene expansion west from South China, and also one individual in Java.

### 3.6.2 Haplogroup E1

Haplogroup E is a primary branch of M9, which divides into E1 and E2, each further subdivided into 2 subclades, E1a, E1b, E2a and E2b (Figure 3.20). As mentioned earlier, all 4 subclades have a very distinctive geographic distribution in ISEA, and my analyses have further added on relevant dates on the demographic history of the region previously reported by Soares *et al.* (2008). Only E1a is found in Taiwan, while E1b is largely restricted to ISEA, with occasionally extending to New Guinea (E1a2), Peninsular Malaysia and Thailand, indicating that both subclades arose in ISEA and dispersed fairly recently east and west. The age estimation in Soares *et al.* (2008) was calculated using coding region and uncorrected mutation rate, which tended to be overestimated; here my results (using corrected molecular clock by Soares *et al.*, 2009) should give more realistic dates and is useful for checking the conclusions there.

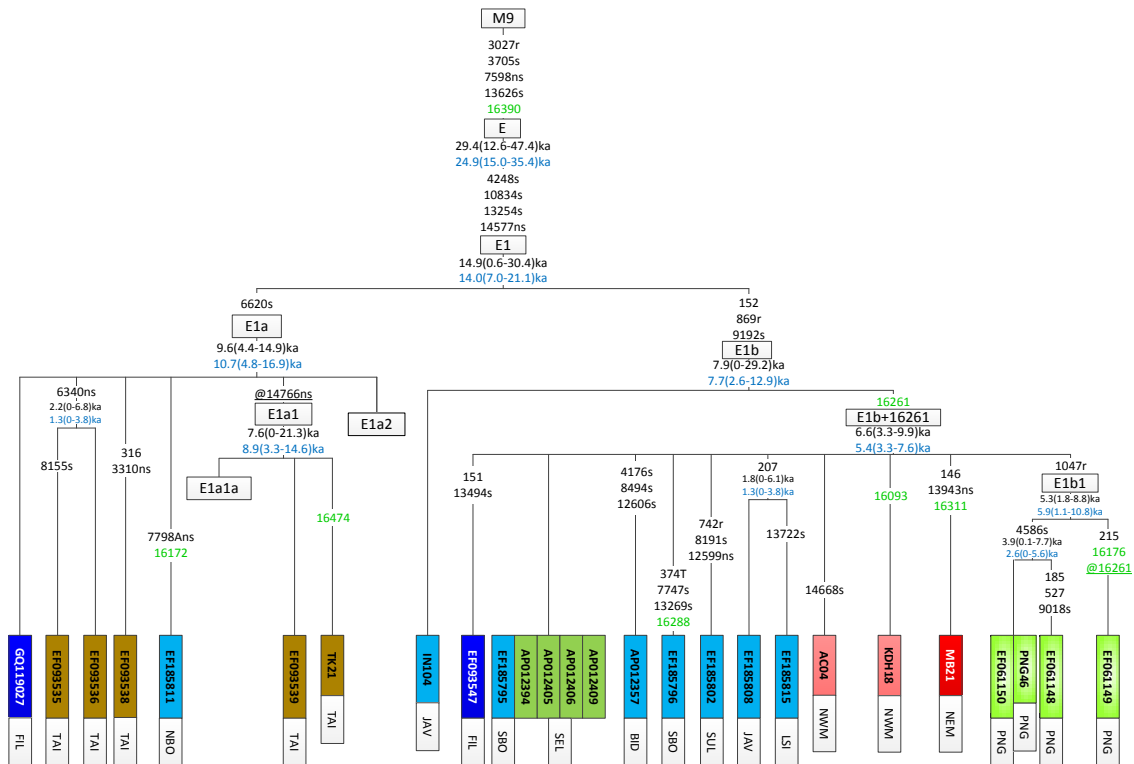
**E1** dates to ~15 ka, and is found only very rarely in China (one each from Guangxi, Xinjiang and Guangdong) and **E1a** is entirely absent in China; the HVS-I data confirms this result (Hill, 2005). In Figure 3.23, E1a dates to ~10 ka (revised from 12 ka in Soares *et al.*, 2008); it is most commonly found in the northern part of ISEA in the Philippines (Tabbada *et al.*, 2008) and North Borneo, and also in Taiwan (Soares *et al.*, 2008). **E1a1** dates to ~8 ka (formerly 9 ka) with two basal lineages seen only in Taiwan. **E1a1a** dates to ~8 ka (previously ~10 ka), and is seen in Taiwan (Soares *et al.*, 2008), the Philippines (Gunnarsdóttir *et al.*, 2011a), South Borneo (Soares *et al.*, 2008), Sumatra (Gunnarsdóttir *et al.*, 2011b) and the North- and Southeast Peninsular Malay (this study). There are at least four subclades nested within E1a1a, including E1a1a1 and three other unnamed subclades. A subclade defined by transitions at nps 131 and 8577, dating to ~4 ka, is seen in Taiwan and the Philippines (Soares *et al.*, 2008). The second subclade defined by a transition at np 709, dating to ~5 ka, is seen in Taiwan (Soares *et al.*, 2008) and Sumatra (Gunnarsdóttir *et al.*, 2011b). Lastly, the third subclade defined by a transition at np 9699, dating to ~4 ka, is seen in South Borneo (Soares *et al.*, 2008) and Southeast Peninsular Malay (this study).

In E1a1a, Taiwanese and ISEA lineages are interleaved within it so that a direction of dispersal is less clear. Soares *et al.* (2008) showed that the overall haplotype diversity (translated into an age estimate) is considerably higher (older ~13 ka) in ISEA than in Taiwan (younger ~11 ka). My results echoed their finding that E1a2, one of the two major subclades



of E1a, is only present in ISEA (at an older age of ~8 ka) compared to E1a1 (~7.6 ka), confirming an origin of E1a in ISEA which then spread into Taiwan mid-Holocene.

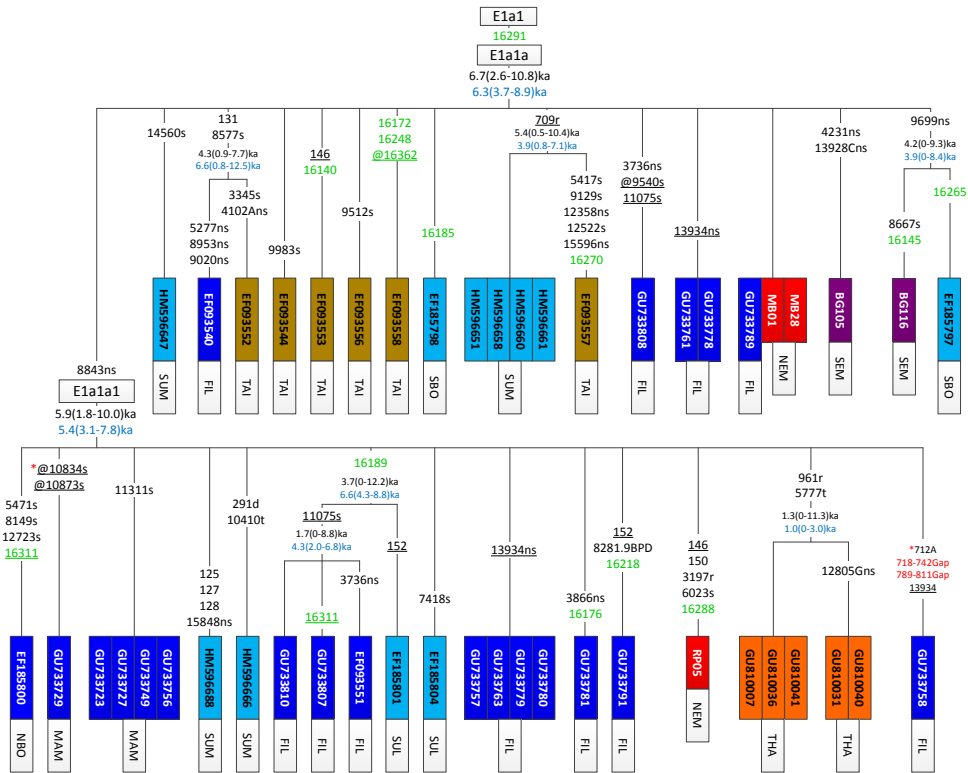
**E1a1a1** dates to ~6 ka, and is spread across ISEA/Peninsular Malaysia (Figure 3.24). It is seen widely in the Philippines including the negrito Mamanwa (Gunnarsdóttir *et al.*, 2011a), Sulawesi (Soares *et al.*, 2008) and Sumatra (Gunnarsdóttir *et al.*, 2011b), a Northeast Peninsular Malay (this study) and five individuals in Thailand (Pradutkanchana, Ishida and Kimura, 2010). There are two Southeast Asian subclades within E1a1a1; the first has a transition at np 16189, dating to ~4 ka, which splits between Sulawesi and the Filipinos from Surigaonons and Visayan Island around ~2 ka. The second subclade is defined by transitions at nps 961 and 5777 and dates to ~1 ka, and is found only in Thailand (Pradutkanchana, Ishida and Kimura, 2010), indicating a very recent founder effect in Thailand.



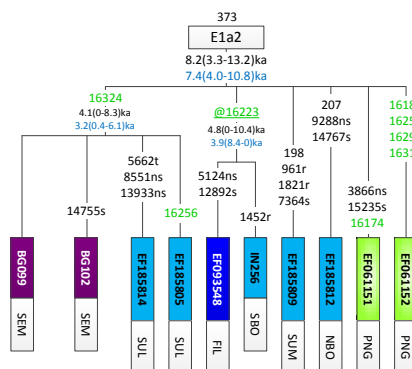
**Figure 3.23** The phylogeny of haplogroup E1 excluding E1a1a and E1a2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BID – Bidayuh Sarawak, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NBO – North Borneo, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SBO – South Borneo, SUL – Sulawesi, PNG – Papua New Guinea, TAI – Taiwan)

**E1b** dates to ~8 ka and the basal lineage is seen in Java (Archaeogenetics Research Group, Huddersfield) while the rest of the lineages are nested within a subclade further defined by a transition at np 16261, dating to ~7 ka (Figure 3.23). This subclade is widely distributed in ISEA, seen in Aboriginal Malay Seletar (Jinam *et al.*, 2012), Peninsular Malay

(this study), Indonesia (Soares *et al.*, 2008; Jinam *et al.*, 2012), the Philippines (Soares *et al.*, 2008), and further nested within is subclade **E1b1** found in Papua New Guinea (Friedlaender *et al.*, 2007; Archaeogenetics Research Group, Huddersfield) dating to ~5 ka. The whole-mtDNA tree indicates E1b likely to have been originated in ISEA during mid-Holocene and dispersed across region and as west as Oceania.



**Figure 3.24** The phylogeny of haplogroup E1a1a. Sequence marked by “\*” are excluded from age estimations. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, MAM – Philippines Mamanwa, NBO – North Borneo, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, PNG – Papua New Guinea, TAI – Taiwan)



**Figure 3.25** The phylogeny of haplogroup E1a2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, NBO – North Borneo, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUL – Sulawesi, PNG – Papua New Guinea)

**E1a2** dates to ~8 ka (formerly ~9 ka in Soares *et al.*, 2008, Figure 3.25). It has an extensive distribution in Indonesia (Soares *et al.*, 2008), the Philippines and North Borneo (Soares *et al.*, 2008), also spreading to Peninsular Malaysia and as far east as Papua New Guinea (Friedlaender *et al.*, 2007), despite the low number of E1a2 sequences reported. A subclade with np 16324, dating to ~4 ka, is seen in Sulawesi (Soares *et al.*, 2008), and Southeast Peninsular Malay (this study). A second subclade has a loss at np 16223, dating to ~5 ka and is seen in the Philippines (Soares *et al.*, 2008) and South Borneo (Archaeogenetics Research Group, Huddersfield).

We now turn to the HVS-I data in Figure 3.26 adapted from Hill (2005), which corresponds to the whole-mtDNA trees showing E1a and E1b have a very distinctive geographic distribution where these early Holocene haplogroups are largely restricted to ISEA. The most common E1\* types are found in the Taiwanese Aborigines, ISEA and Peninsular Malaysia. The E1\* gives rise to another clade called E1b with a transition at np 16261. E1b types are mainly found in ISEA, particularly Sulawesi, and some individuals from Peninsular Malaysia.

Similar to haplogroup E1b, the HVS-I data of haplogroup E1a1a (recognisable in the HVS-I data) is most common in Eastern Indonesia, especially Sulawesi and northern Borneo (Figure 3.27). It is also found across the rest of ISEA, including Sumatra and the Philippines, at much lower levels. The root type is the most common type and mainly found throughout the region, but most commonly in Sulawesi and Taiwan. The others are mostly one-step derivatives, one type with transition at np 16185 is particularly prevalent in South Borneo (represented by a single sample in the complete mtDNA tree, Soares *et al.*, 2008).

The new analysis has also shown some sink recipients in Thailand (seen in subclade E1a1a1) and PNG (in E1a2 and E1b1). E1a1a1 is more diversified in ISEA and virtually absent in Taiwan, with a fairly recent migration into Thailand (possibly through Peninsular Malaysia). Again, strongly suggesting E1a and E1b both arose in ISEA and dispersed recently east and west.

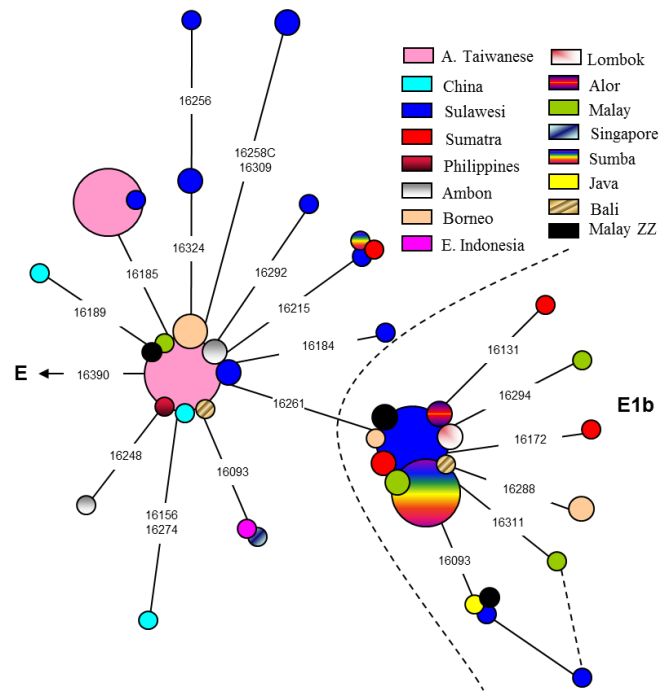


Figure 3.26 HVS-I network of E1\* and E1b. Figure adapted from Hill (2005).

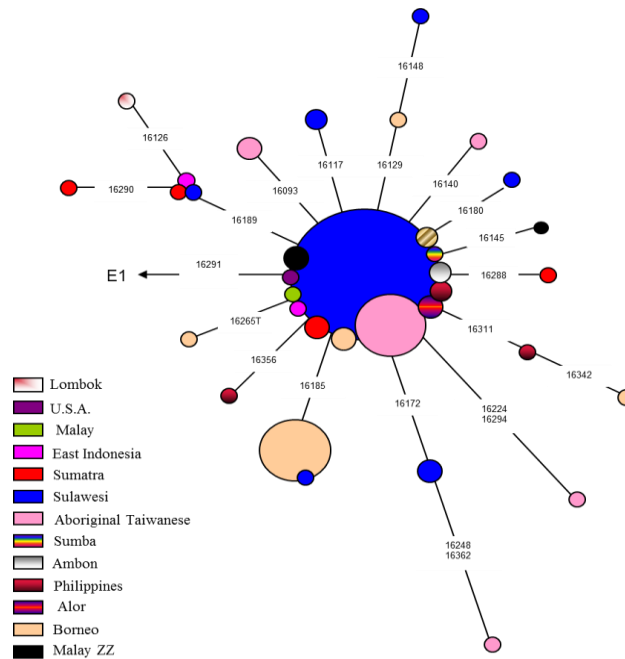


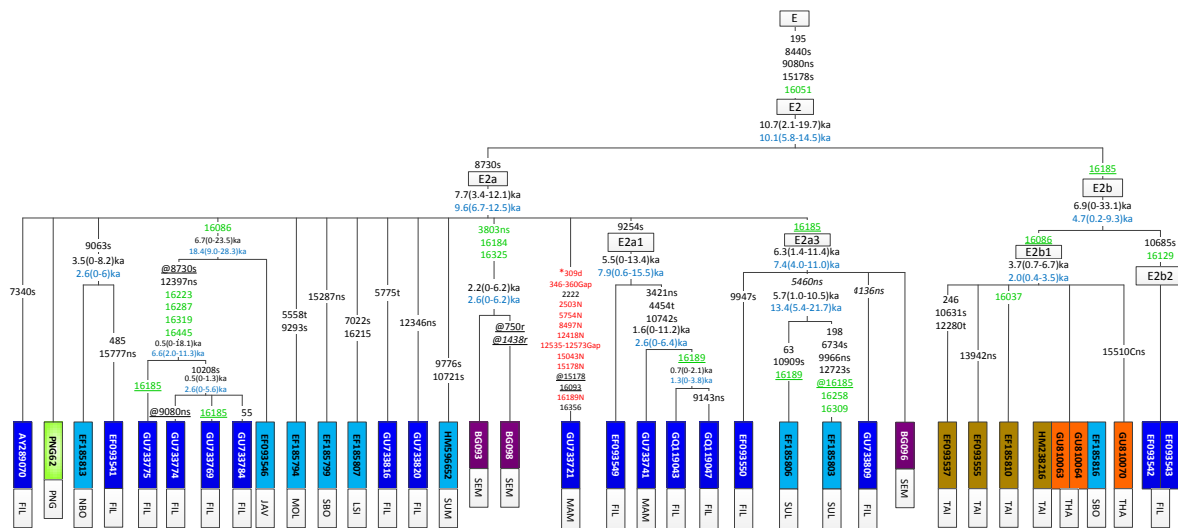
Figure 3.27 HVS-I network of E1a1a. Figure adapted from Hill (2005).

### 3.6.3 Haplogroup E2

**E2** dates to ~11 ka (previously ~9.5 ka in Soares *et al.*, 2008), and it is divided into E2a and E2b (Figure 3.28). **E2a** dates to ~8 ka (formerly ~6.7 ka in Soares *et al.*, 2008), widely seen in the Philippines, Eastern Malaysia/Indonesia (Ingman and Gyllensten, 2003; Soares *et al.*, 2008; Gunnarsdóttir *et al.*, 2011a), Peninsular Malay (this study), and Papua New Guinea

(Archaeogenetics Research Group, Huddersfield). There are four subclades nested within E2a: E2a1, E2a3, and three other unnamed subclades. **E2a1**, dating to ~5.5 ka, restricted to the Philippines (Soares *et al.*, 2008) with an instance of the Philippines negrito Mamanwa (Gunnarsdóttir *et al.*, 2011a) nested within the Filipino subclade (Tabbada *et al.*, 2010). **E2a3** dates to ~6 ka, and is seen in the Philippines (Soares *et al.*, 2008; Gunnarsdóttir *et al.*, 2011a) and Southeast Peninsular Malaysia (this study), with a subclade nested within (~6 ka) formed by lineages from Sulawesi (Soares *et al.*, 2008).

E2a includes three other nested subclades. One is found in North Borneo and the Philippines (Soares *et al.*, 2008); the second in Java and Manobo Filipinos (Soares *et al.*, 2008; Gunnarsdóttir *et al.*, 2011a). Lastly, a subclade dating to ~2 ka, includes two Southeast Peninsular Malay, suggesting an arrival in the Peninsula from ISEA by about that time.



**Figure 3.28** The tree of haplogroup E2. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. GU733721 is excluded from age estimations. (FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, MOL – Moluccas, Indonesia, NBO – North Borneo, PNG – Papua New Guinea, SBO – South Borneo, SEM – Southeast Peninsular Malay, SUM – Sumatra)

**E2b** dates to ~7 ka (previously ~4.3 ka in Soares *et al.*, 2008) and divides into E2b1 and E2b2. **E2b1** dates to ~4 ka, and is reported in Taiwan (Soares *et al.*, 2008; Loo *et al.*, 2011), Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Palangkaraya in southern Borneo, Indonesia (Soares *et al.*, 2008); one sequence (unusually, the root type of E2b1) is shared across all three of these regions. **E2b2** is extremely rare and has been only seen in two individuals from the Philippines with a single identical sequence (Soares *et al.*, 2008).

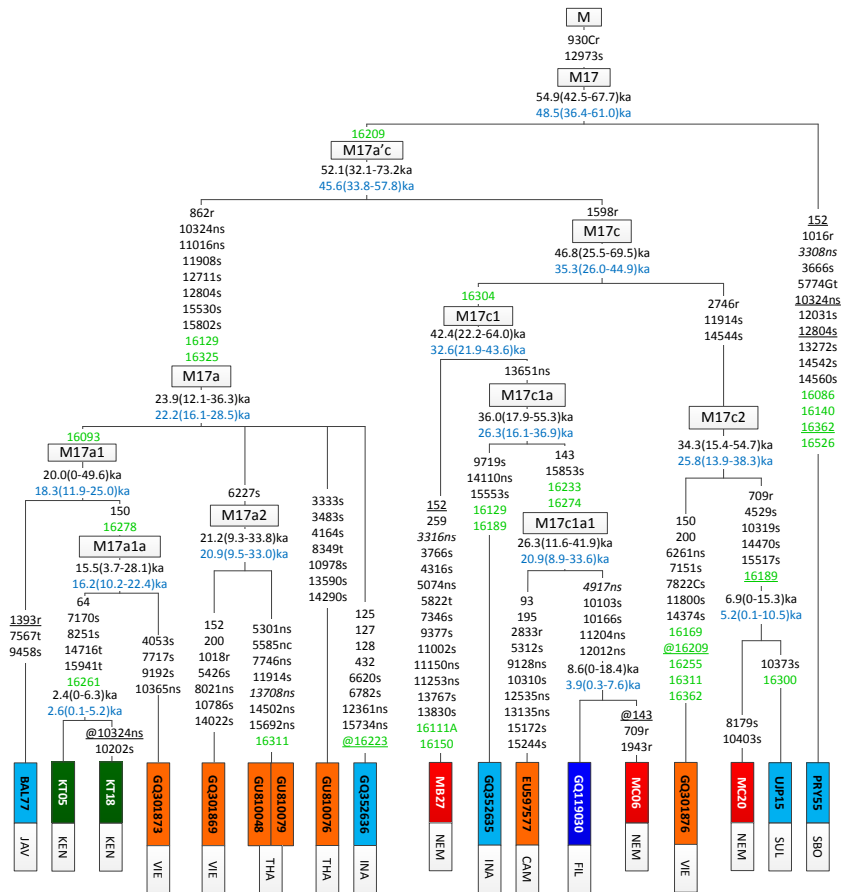
Again, similar signals to E1a1 and E1a2 are seen in E2a and E2b. E2a dates older (~8 ka) compared to E2b (~7 ka), and Taiwan (in this case, although, as well as Thailand) are

only seen within E2b1 and not E2a, indicating an origin in ISEA. Indeed, Soares *et al.* (2008) found that prior to the proposed Austronesian expansions from China/Taiwan, haplogroup E shows pattern of postglacial expansion and dispersal (on the basis of HVS-I data in Hill *et al.*, 2007). They suggested the spread of these genetic signatures is the impact on coastal-dwelling populations of the rapid global warming and sea-level rises that led to the inundation of the Sunda shelf by meltwater at the end of the last Ice Age (Oppenheimer, 1998; Lin *et al.*, 2005). Sea levels began to rise gradually after the end of the LGM ~19 ka, probably in three major episodes of flooding due to ice sheet collapse. These events are sometimes referred to as Catastrophic Rise Events 1-3, at ~14.5 ka, 11.5 ka and probably also ~7.5 ka (Blanchon and Shaw, 1995; Oppenheimer, 1998; Pelejero *et al.*, 1999; Hanebuth *et al.*, 2000; Voris, 2000; Lambeck and Chappell, 2001; Bird *et al.*, 2005). It has been suggested that lineage expansions throughout Island SEA tend to coincide with the three rapid sea level rises and were preceded by long lineage lines indicating immediate loss as the sea level rose. These flooding episodes would have triggered major displacements of human groups living on the Sunda coastline and had an important role in shaping subsequent life in the region especially its maritime orientation and the development of sailing technology (Oppenheimer, 1998; Solheim, 2006).

### 3.7 Haplogroup M17

Figure 3.29 is the phylogeny of haplogroup M17, a very rare and ancient Sunda-specific lineage, with a widespread relict distribution. Here I propose that M17 is re-defined by two polymorphisms at nps 930C and 12973 (excluding np 16209 as suggested by Phylotree) to include a basal lineage detected in Palangkaraya of South Borneo (Archaeogenetics Research Group, Huddersfield), dating to ~55 ka. **M17a'c** (newly named) is found mainly in MSEA/Malay Peninsula, though with isolated instances throughout Indonesia and the Philippines, and divides into subclades M17a and M17c at ~52 ka. **M17a** has undergone high drift resulting in a date of ~24 ka. Closely related M17a lineages are seen in the Kintak and Kensui Semang (Malay Peninsula), in Indonesia (Archaeogenetics Research Group, Huddersfield; Tabbada *et al.*, 2010), in Thailand (Pradutkanchana, Ishida and Kimura, 2010), and in Vietnam (Peng *et al.*, 2010). **M17a1** dates to ~20 ka and a basal lineage is detected in Java (Archaeogenetics Research Group, Huddersfield). **M17a1a**, dating to ~16 ka, is found in MSEA: in Vietnam (Peng *et al.*, 2010) and with Semang Kensiu nested in a subclade that

dates to ~2 ka. M17a2, dating to ~21 ka, is again confined to MSEA, being seen in Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Vietnam (Peng *et al.*, 2010). M17a expands around the end of the LGM on the Sunda shelf, with an offshoot arrives in the Semang who share the MRCA with Vietnam during the Pleistocene.



**Figure 3.29** The tree of haplogroup M17. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CAM – Cambodia, FIL – Philippines, INA – Indonesia, JAV – Java, Indonesia, KEN – Semang Kensi, NEM – Northeast Peninsular Malay, SBO – South Borneo, SUL – Sulawesi, THA – Thailand, VIE – Vietnam)

Haplogroup **M17c** has rare, isolated ‘Sunda’ occurrences throughout MSEA and ISEA dividing, ~47 ka, into M17c1 and M17c2. **M17c1** dates to ~42 ka with a basal branch in a Northeast Peninsular Malay (this study). **M17c1a**, dating to ~36 ka, with a basal haplotype in Indonesia (Tabbada *et al.*, 2010), with a further subclade **M17c1a1**, seen in Cambodia (Hartmann *et al.*, 2009). Another subclade nested within, dating to ~9 ka, is again widespread in SEA being seen in a Northeast Peninsular Malay (this study) and a Filipino (Tabbada *et al.*, 2010).

M17c2 dates to ~34 ka, and is seen in Vietnam (Peng *et al.*, 2010). A subclade below dates to ~7 ka, again both from MSEA, in a Northeast Peninsular Malay, and ISEA in Sulawesi, Indonesia (Archaeogenetics Research Group, Huddersfield).

The HVS-I data indicated that M17a (previously known as M\*) is found only in the Semang Kensiu and not elsewhere, emphasising its extreme rarity. The whole-mtDNA tree indicates that at least the M17a1a in Kensiu has a deep Sunda Origin, possibly centred in MSEA.

### 3.8 Haplogroup M12'G

M12'G is defined by a transition at np 14569, dividing into M12 and G ~60 ka (Figure 3.30). The phylogeny of M12'G includes 104 complete sequences: 79 G and 25 M12. 48 of these came from Japan (Tanaka *et al.*, 2004) and 23 from China (Kong *et al.*, 2003a; Kong *et al.*, 2006; Kong *et al.*, 2011; Peng *et al.*, 2011b; Zheng *et al.*, 2011). The large number of Japanese complete sequences compared to Chinese potentially over-represented on the trees but they are highly diversified and localised in Japan. The HVS-I data pooled from several papers (of different sample size) by Tanaka *et al.* (2004) has shown that haplogroup G is found at high frequencies in the Japanese (6.86%), indigenous Ainu (4%) and Ryukyuan (9.8%), and Koreans (5.77%), followed by the Chinese and Central Asian at lower frequency (Table 2 in Tanaka *et al.*, 2004).

Haplogroup G includes subclades G1, G2, G3 and G4. Haplogroup G was initially thought to be a basal branch of haplogroup M, but it has now been subsumed by haplogroup M12'G which includes all the branches of haplogroup G and M12, the former with a broadly East Asian distribution (suggesting a likely source in China) and the latter found across Southeast Asia (mainly MSEA) and South China and more likely originating in MSEA. Haplogroup G and most of its subclades appear to have an origin and spread limited to China and significant spread to Japan, with several exceptions. These exceptions are reported at low levels: G1b is seen in a North Asian Eskimo, and G1c is reported in China as well as an Aboriginal Malay (Figure 3.30). G2a1, G2b and G3b all have representatives in India.

Haplogroup M12 is divided into M12a and M12b. M12a may have an origin in MSEA/South China (there is a deep basal lineage in Northeast Malaysia: Figure 3.30)



Subclade M12a1b1 is reported in East Asia, South Asia and Northeast Peninsular Malay. On the other hand, M12b1 is found in SEA, while M12b2 in India.

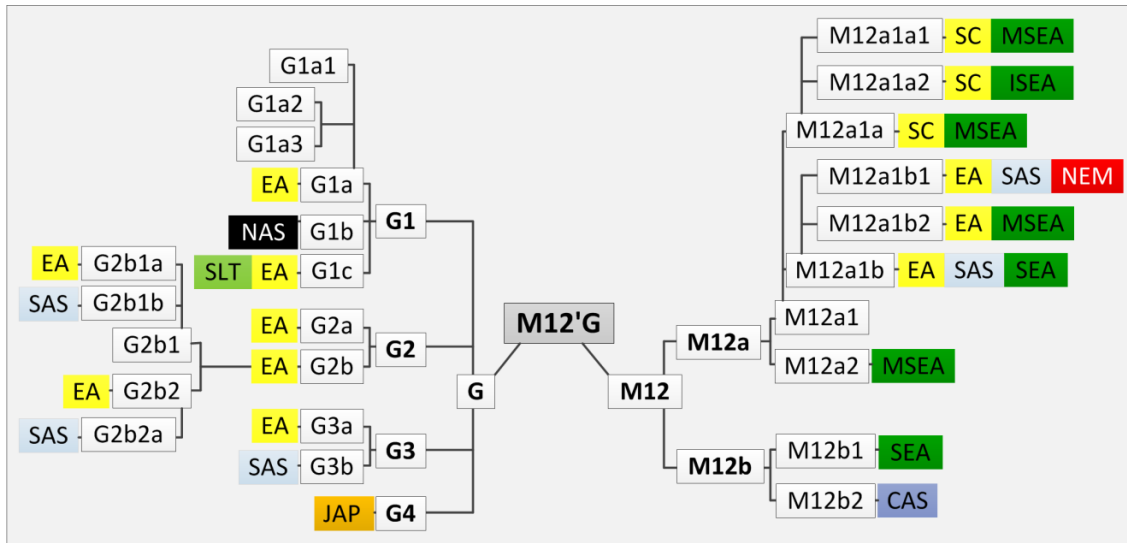
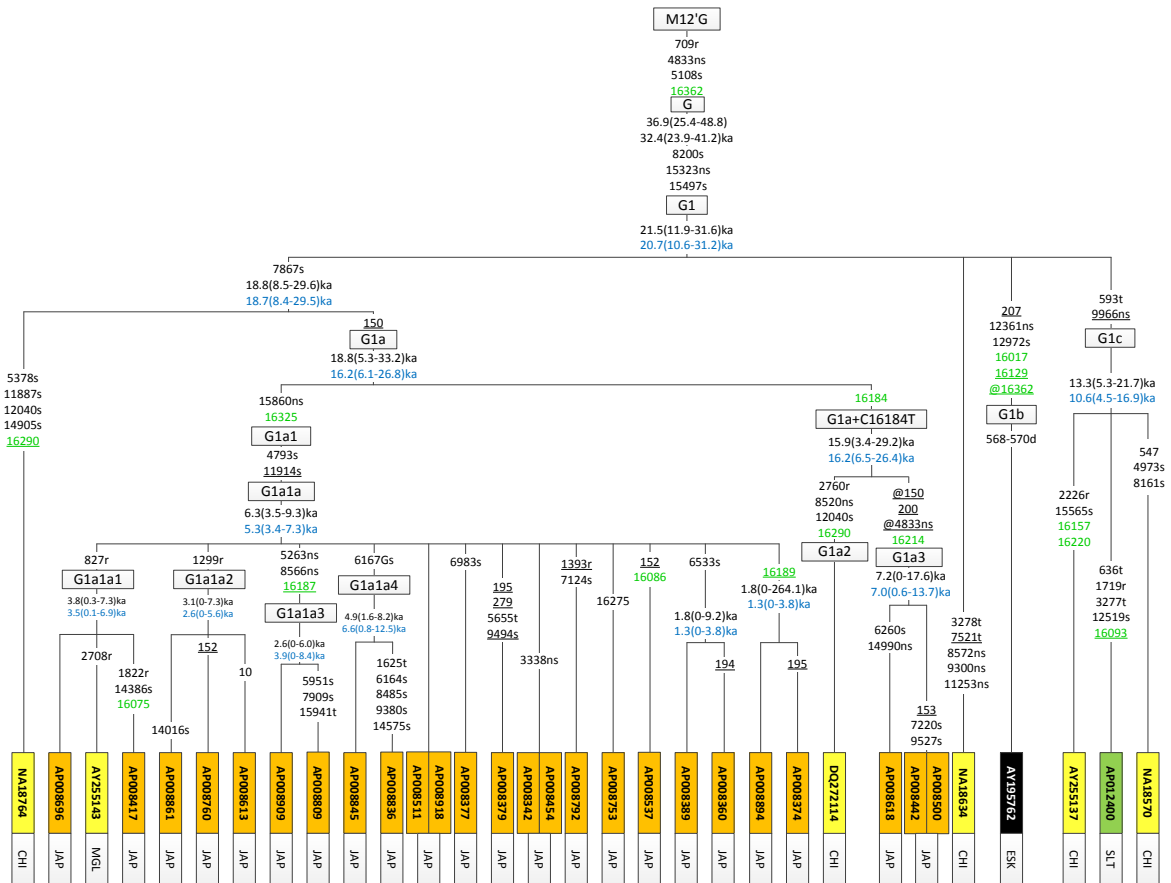


Figure 3.30 Schematic diagram of haplogroup M12'G and its major subclades distribution. (CAS – Central Asia, EA – East Asia, ISEA – Island Southeast Asia, JAP – Japan, MSEA – Mainland Southeast Asia, NAS – North Asia, NEM – Northeast Peninsular Malay, SAS – South Asia, SC – South China, SEA – Southeast Asia, SLT – Aboriginal Malay Seletar)

### 3.8.1 Haplogroup G1

**G1** dates to the LGM ~22 ka and is divided into G1a, G1b and G1c (Figure 3.31). A lineage from Northern China (Zheng *et al.*, 2011) is likely to show a reversion at np 150, which appears pre-G1a-like and dates to ~19 ka. **G1a** dates to ~19 ka and is commonly found in Japan and China (Kong *et al.*, 2003a; Tanaka *et al.*, 2004). G1a can be divided into two subclades: G1a1a and G1a+16189 (including G1a2 and G1a3). **G1a1a** dates to ~6 ka, and seen mainly from Japan (Tanaka *et al.*, 2004). It has at least six subclades including G1a1a1, G1a1a2, G1a1a3, G1a1a4, and two other unnamed subclades. **G1a1a1**, dating to ~4 ka, seen in Japan and Inner Mongolia (Kong *et al.*, 2003a). The rest of the subclades are only found in Japan: **G1a1a2** dates to ~3 ka, **G1a1a3** ~ 3 ka and **G1a1a4** ~5 ka, as well as the two unnamed subclades. Subclade defined by a transition at np 6533 dates to ~2 ka, and subclade defined by a transition at np 16189, also ~2 ka.

**G1a2** and **G1a3** shared the node G1a+C16184T, which dates to ~16 ka. G1a2 is represented here by one Chinese sample (Kong *et al.*, 2006). G1a3 is a rare haplogroup estimated at ~7 ka, and is found to be localised in Japan (Tanaka *et al.*, 2004). Both G1a1a and G1a3 could potentially be Neolithic dispersals into Japan/Korea.



**Figure 3.31** The tree of haplogroup G1. Time estimates shown for clades are ML (in black) and averaged distance ( $p$ ; in blue) in ka. (CHI – China, ESK – Eskimo, JAP – Japan, MGL – Inner Mongolia, China, SLT – Aboriginal Malay Seletar)

**G1b** is represented here by one lineage found in the Kamchatka Peninsula, Russia by Mishmar *et al.* (2003). **G1c** is a rare haplogroup, dating to ~13 ka, found at low levels in northeast China (Kong *et al.*, 2003a; Zheng *et al.*, 2011) and, interestingly, in the Seletar, Aboriginal Malay from the southern tip of the Malay Peninsula (Jinam *et al.*, 2012), which indicates a northern/China origin for some of the Malay lineages. Besides, a complete mtDNA sequence of haplogroup G1c (not in the tree) is also reported in Korea (Derenko *et al.*, 2007).

### 3.8.2 Haplogroup G2

**G2**, dating to ~31.5 ka, and is found throughout China (Kong *et al.*, 2006; Kong *et al.*, 2011; Zheng *et al.*, 2011), Japan (Ingman *et al.*, 2000; Tanaka *et al.*, 2004; Nohira *et al.*, 2010), and at lower levels in India and Pakistan (Kong *et al.*, 2006; Chandrasekar *et al.*, 2009). Detailed description is available in Appendix E.

However, the HVS-I data shows that G2 is relatively abundant in northern China and central Asia, reaching considerable high frequencies in the Mansi and Tuvinians in Southern Siberia (Tanaka *et al.*, 2004). Subclades G2b1 and G2b2 indeed indicate here of deep ancestry in China before the LGM and very early expansions from East Asia into India, which is otherwise undetected by looking only at the HVS data because of the lack of HVS-I motifs.

### 3.8.3 Haplogroups G3 and G4

**G3** occurred at lower levels in Japan (Tanaka *et al.*, 2004) and China (Kong *et al.*, 2006). **G4**, dating to only ~3 ka, is a very rare haplogroup – it is only detected in two Japanese sequences reported by Tanaka *et al.* (2004). Detailed description is available in Appendix E.

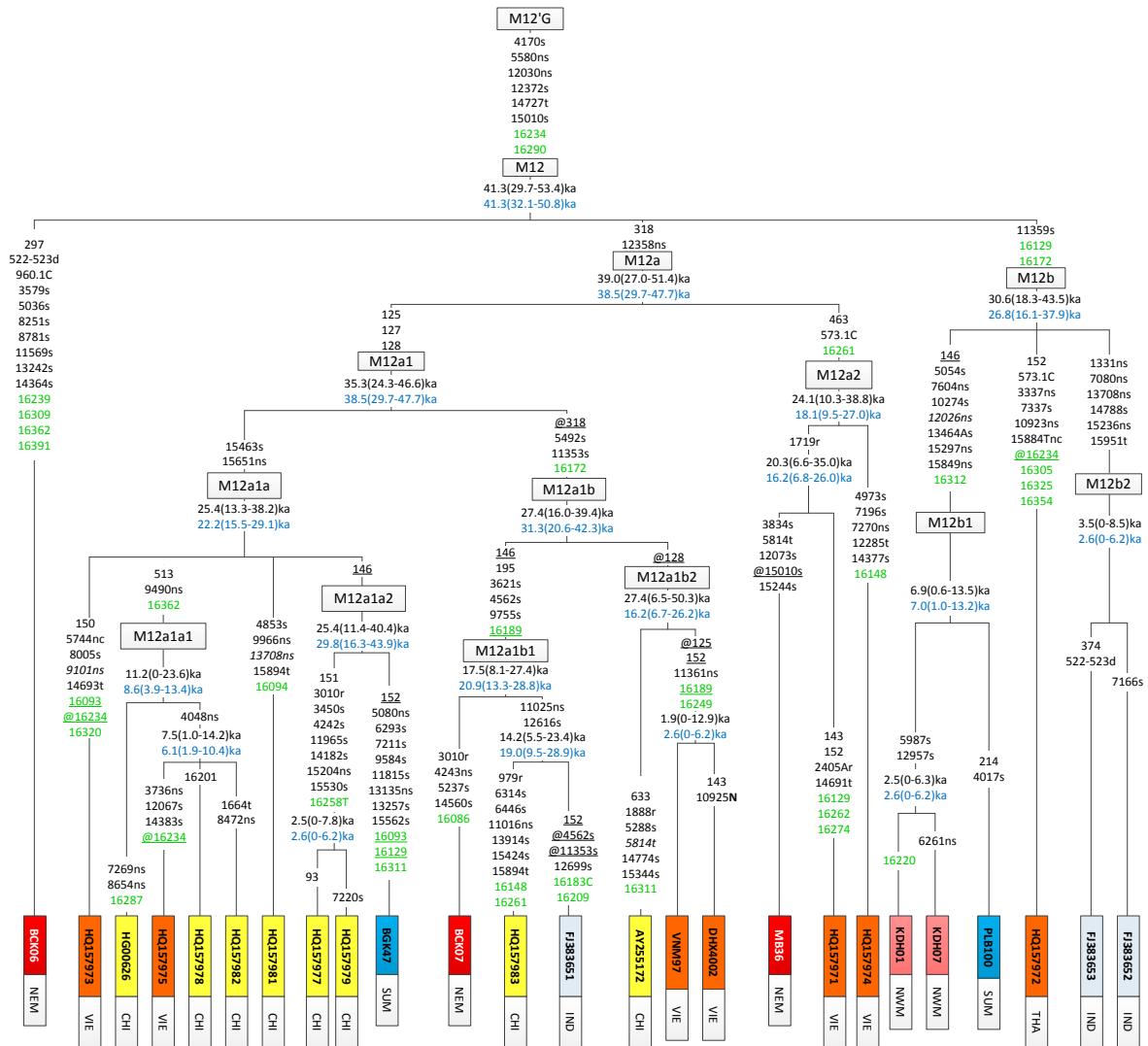
Haplogroup G and its major subclades pre-date the LGM and are widely distributed in northern and eastern China, Japan, South Asia including India and Pakistan, with several single instances from Russia (in G1b), the Aboriginal Malay Seletar (in G1c) and Georgia (in G2a1). Apart from the G1c (~13 ka) Aboriginal Malay Seletar lineage from Peninsular Malaysia (Jinam *et al.*, 2012), lineages of haplogroup G are virtually absent in SEA and the Pacific. This brings to mind the traditional *Orang Asli* “layer-cake” theory (Cole, 1945; Carey, 1976; Birdsell, 1993) that suggested the Aboriginal Malays arrived in the final wave into Peninsular Malaysia (as the influx of Mongoloids together with the colonisation of the Indo-Malaysian Archipelago by the “Proto-Malays”), G1c lineage in Seletar has shown a longstanding relationship with the Han Chinese in northern China which indicates at least a small fraction of the Seletar did not come from ISEA, although it is only indicated by a singleton from Seletar.

### 3.8.4 Haplogroup M12

**M12** dates to ~41 ka and is divided into M12a and M12b. There are three basal branches: M12a, M12b and a paraphyletic lineage seen a single Northeast Peninsular Malay (this study). The overall distribution suggests long-term ancestry in west Sunda/South China, with a (possibly recent) offshoot to India in the form of M12b2 (which the HVS-I database suggests it is also present in South China and Thailand).

**M12a** dates to ~39 ka and is divided into M12a1 (~35 ka) and M12a2 (~24 ka) (Figure 3.32). **M12a1a** dates to the LGM ~25 ka and the origin appears to lie in northern MSEA as

the basal lineages are found in Vietnam and South China (Peng *et al.*, 2011b; Zheng *et al.*, 2011). **M12a1a1**, dating to ~11 ka, is found in South China and Vietnam. **M12a1a2** dates to ~25 ka and is found in Sumatra (Archaeogenetics Research Group, Huddersfield) and South China (Peng *et al.*, 2011b).



**Figure 3.32** The tree of haplogroup M12. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, IND – India, NEM – Northeast Peninsular Malay, NWM - Northwest Peninsular Malay, SUM – Sumatra, THA – Thailand, VIE – Vietnam)

Similar to M12a1a, **M12a1b** appears to have a pre-LGM origin in MSEA or South China (Figure 3.32). **M12a1b1** (dates to ~17.5 ka) is seen in Northeast Peninsular Malaysia (this study), and nested within a subclade found in Yunnan, China (Peng *et al.*, 2011b) and Gallong, India (Chandrasekar *et al.*, 2009) ~14 ka. **M12a1b2** (~27 ka) is seen in Guangdong China (Kong *et al.*, 2003a) and nested within a small clade (~2 ka) in Vietnam

(Archaeogenetics Research Group, Huddersfield). **M12a2** dates to ~24 ka and is seen exclusively in MSEA with a basal branch in Vietnam, and ~20 ka with derivative haplotypes in Vietnam (Peng *et al.*, 2011b) and a Northeast Peninsular Malay (this study).

**M12b** dates to ~31 ka and the basal lineage is seen in Northern Thailand (Peng *et al.*, 2011b). **M12b1** dates to ~7 ka and seen in Sumatra (Archaeogenetics Research Group, Huddersfield), and two nested clusters, one (~3 ka) represented in two Northwest Peninsular Malay and a Sumatran. The other, **M12b2**, dating to ~4 ka and is seen only in India (Chandrasekar *et al.*, 2009).

We can now turn to the HVS-I data, where the most derived types are seen in ISEA but MSEA has the highest levels of M12 types, with frequencies rise up to ~8 % in the southeastern parts of the region, ~2 % in northeastern MSEA, ~3 % in Malaysia, and ~1 % of the sample in Coastal China (frequency gradient distribution of M12 in Mormina, 2007). Considering the frequency gradient, the HVS-I data described above, and the phylogeographic distribution of M12, it is likely that haplogroup M12 have a pre-LGM origin in MSEA and a coastal distribution in MSEA, ISEA, and South China. Between the period of LGM and late Pleistocene, MSEA and ISEA would have joined as a single landmass, Sundaland, allowing M12 to spread along the northeastern coast, hence the extant relict descendants are preserved in the Peninsular Malay, Sumatra and India.

### 3.9 Haplogroup M29'Q

Haplogroup **M29'Q** dates to ~61 ka and is divided into M29 and Q (Figure 3.33). M29 (or rather **M29b**) is represented in the tree by an instance from Papua New Guinea (Friedlaender *et al.*, 2007). Haplogroup **Q** is an Oceanian haplogroup most commonly reported in Papua New Guinea and West Papua (Redd *et al.*, 1995; Lum *et al.*, 1998; Archaeogenetics Research Group, Huddersfield), and at lower levels in Vanuatu, Polynesia and Micronesia (Sykes *et al.*, 1995; Lum *et al.*, 1998; Hagelberg *et al.*, 1999). The phylogeny of M29'Q includes 20 Q complete sequences, and one belonged to M29b.

Haplogroup **Q** dates to ~46 ka, and can be divided into Q1'2 and Q3. **Q1** dates to the LGM, ~24 ka, and includes subclades **Q1+@T16233C** (~24 ka) and an unnamed subclade (~1 ka), seen in PNG, the Cook Islands, Samoa, Vanuatu, Bougainville, two lineages from Peninsular Malaysia and one from the Philippines. **Q1+@T16223C** includes Q1a and Q1b

**Q3** dates to ~39 ka and commonly found in PNG, with a basal lineage found in Northwest Peninsular Malaysia (this study). **Q3a** dates to ~33 ka, and subclades nested within date to ~18 ka and ~5 ka, and are restricted to PNG (Ingman and Gyllensten, 2003;

Hartmann *et al.*, 2009). Q3b is represented here by an instance from PNG (Friedlaender *et al.*, 2007).

We can now complement the whole-mtDNA tree with HVS-I data. The HVS-I network of haplogroup Q1 (Figure 3.34) shows a number of distinct lineages sampled in ISEA. It is found at low levels in Banjarmasin and Kota Kinabalu of Borneo, Bali, Manado, Toraja, Ujung Padang and Sumba of Indonesia, and Peninsular Malaysia (Hill, 2005; this study). Similarly to haplogroup P, the presence of haplogroup Q in ISEA suggests very recent Holocene gene flow across into ISEA as far as the Malay Peninsula from Near Oceania (as some ISEA lineages were also spreading the other way, e.g. haplogroup E).

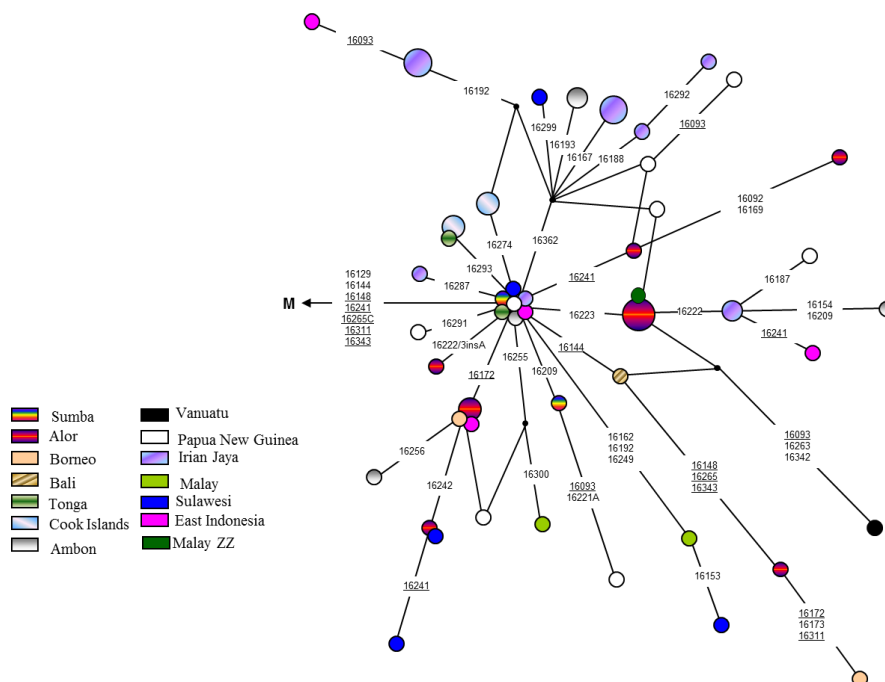
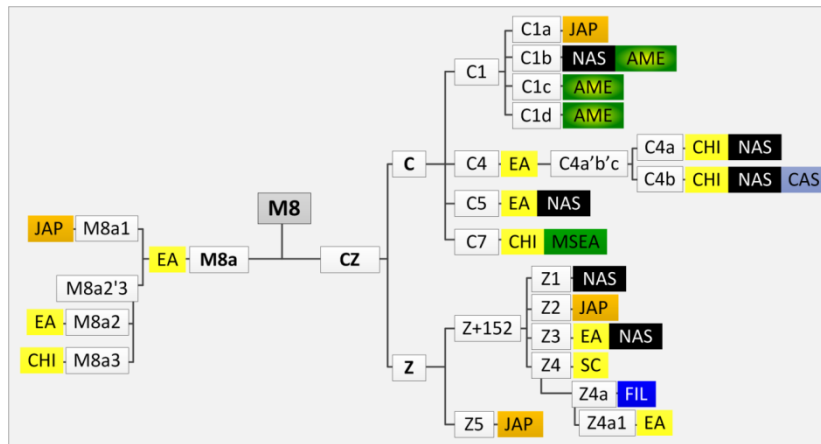


Figure 3.34 HVS-I network of Q1. Figure adapted from Hill (2005).

### 3.10 Haplogroup M8

**M8**, dating to ~ 45 ka, is divided into M8a and CZ (Figure 3.35). Haplogroup **M8a** is widely seen in China and Japan, and in one instance from Siberia, Russia, and is virtually absent in Southeast Asia. Haplogroup **C**, dating to ~27 ka, appears to have a northern origin in East Asia, and consists of haplogroups C1, C4, C5 and C7. C1 is mainly found in the Native American, except for subclade C1a seen in the Japanese and Siberian Russian. C4 shows a basal lineage in northern China, and it is widely distributed across China, Russia and west into Kyrgyzstan. (Detailed descriptions for haplogroups M8a, C1 and C4 are available

in Appendix E). Similarly to C4, C5 is seen in Siberian Russian, China and Japan. C7 is seen in northern China, Thailand and Peninsular Malaysia, indicating this is likely a single instance of a long-range late glacial dispersal south. Similar to M8, haplogroup **Z** also appears to have a northern origin in East Asia, and includes subclades Z1, Z2, Z3, Z4 and Z5. Z1 is seen in Russia, Z2 in Japan, and Z3 in East and North Asia. Lastly, Z4 has a basal lineage in South China, its subclade Z4a has a basal lineage in the Philippines, and a further nested subclade Z4a1 is seen in China and Japan. The phylogeny of M8 includes 71 complete sequences: 17 M8, 37 C and 17 Z.



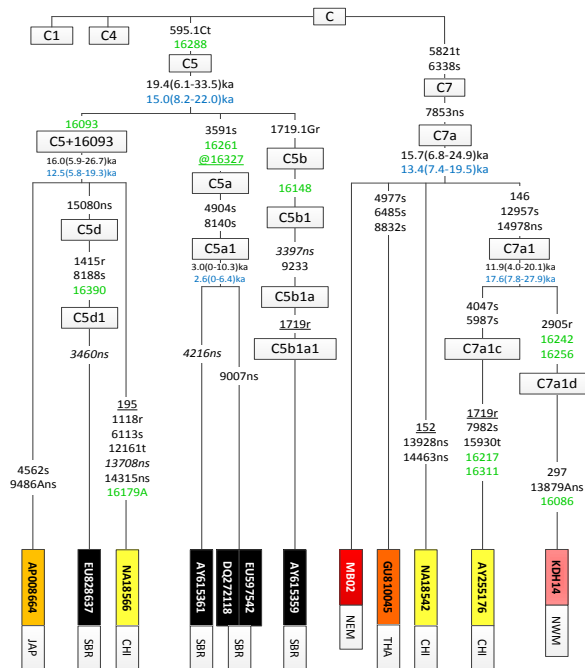
**Figure 3.35** Schematic diagram of haplogroup M8 and the distribution of its major subclades. (AME – America, CAS – Central Asia, CHI – China, EA – East Asia, FIL – Philippines, JAP – Japan, MSEA – Mainland Southeast Asia, NAS – North Asia, SC – South China)

**C5** dates towards the end of the LGM ~19 ka (Figure 3.36). **C5+T16093C** dates to ~16 ka, and is seen in northern China (Zheng *et al.*, 2011), Japan (Tanaka *et al.*, 2004) and Siberia, Russia (Starikovskaya *et al.*, 2005). **C5a1** (~3 ka) is restricted to Siberia (Starikovskaya *et al.*, 2005; Kong *et al.*, 2006; Hartmann *et al.*, 2009) while **C5b1a1** is represented by an instance from Siberia, Russia (Starikovskaya *et al.*, 2005).

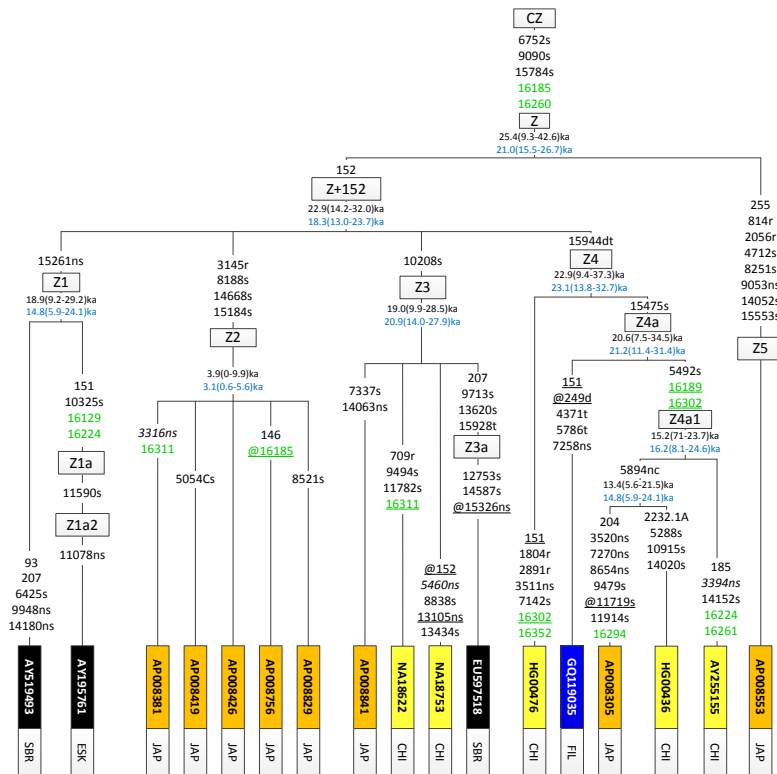
Perhaps the most relevant haplogroup within M8 for this study is haplogroup **C7a**, dating to ~16 ka (Figure 3.36), which has basal lineages in Beijing (Zheng *et al.*, 2011), Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Northeast Peninsular Malaysia (this study). A subclade nested within is **C7a1**, dating to ~12 ka, is shared between instances from Liaoning, China (Kong *et al.*, 2003a) and Northwest Peninsular Malaysia. Haplogroup C as a whole may have a northern origin in East Asia with a divergence time of ~27 ka. The relict descendants in Thailand and Peninsular Malaysia suggest a possible dispersal into MSEA from northern China in an upper bound of 12 ka since there is still a Chinese sample



diverging at this age (seen in C7a1), although it could have been very recent since it is extremely rare in the Malay.



**Figure 3.36** The tree of haplogroups C5 and C7. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NEM – Northeast Peninsular Malay, SBR – Siberian Russia, THA – Thailand)



**Figure 3.37** The tree of haplogroup Z. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, FIL – Philippines, JAP – Japan, SBR – Siberian Russia)

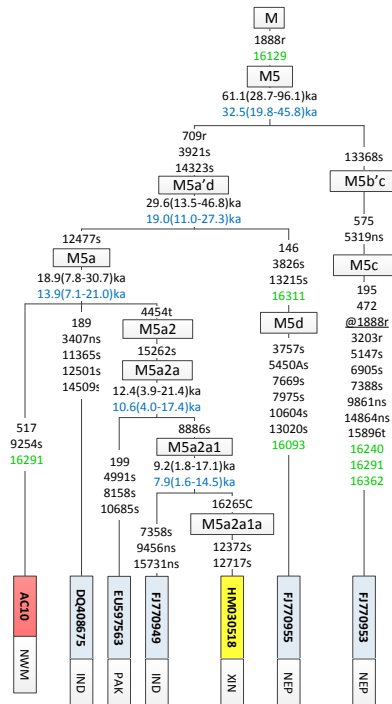
Similar to haplogroup C, haplogroup **Z** looks to have a northern origin in East Asia (Figure 3.37) and is divided into Z+T152C (assuming it is a true clade since np 152 is one of the fastest sites and not likely to define ancient subclades; dates to ~23 ka) and Z5, the latter is represented by a single instance from Japan (Tanaka *et al.*, 2004). **Z1** dates to ~19 ka and is seen in Southern Siberia (Starikovskaya *et al.*, 2005) and Koryak, Eskimo (Mishmar *et al.*, 2003). **Z2** (~4 ka) is restricted to Japan (Tanaka *et al.*, 2004), and **Z3** (~19 ka) is seen in Japan (Tanaka *et al.*, 2004), northern China (Zheng *et al.*, 2011) and northern Siberia (Hartmann *et al.*, 2009). **Z4** dates ~23 ka, where a basal lineage is seen in South China (Zheng *et al.*, 2011), a single instance from the Philippines shared Z4a1, dating to ~21 ka, with lineages from Hubei, Central China (Kong *et al.*, 2003a), South China (Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004) ~13 ka. The Z4 carriers might just have gone across by land at low sea level considering the LGM date into the Philippines, but the most likely scenario could be quite recent given that is a date from one instance.

### 3.11 Haplogroup M4'67

Haplogroup M4'67 dates to ~48 ka, consists of subclades M4, M30 and M37, and two basal lineages from Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Southeast Peninsular Malaysia (this study), all sharing the M4'67 diagnostic site of a transition at np 12007 (Figure 3.38). M4'67 is largely seen in South Asia, especially in West India (Mellars *et al.*, 2013), which suggests an ultimate source there.

**M4** is represented by an instance from West India (Thangaraj *et al.*, 2006). It is also commonly found in the western region of South Asia including Rajasthan, Gujarat, Maharashtra, Pakistan and Punjab (Thangaraj *et al.*, 2006; Chandrasekar *et al.*, 2009; Mellars *et al.*, 2013), also reported in Andhrapradesh in the south and Uttarpradesh of North India (Sun *et al.*, 2006), but not included in this study due to time constraints. **M30** dates to ~26 ka and it is found in Southern India (Ingman and Gyllensten, 2003) and Pakistan (Hartmann *et al.*, 2009), while **M30a1**, dating to ~11 ka, is found in Southern India (Ingman and Gyllensten, 2003) and Northeast Peninsular Malaysia (this study). **M37** dates to ~38 ka, and is found in West India (Thangaraj *et al.*, 2006) and Southwest Peninsular Malaysia (this study). However, it is not possible to infer the origin of these lineages since there are single samples without MSEA specific clade to date.





**Figure 3.39** The tree of haplogroup M5. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (IND – India, NEP – Nepal, NWM – Northwest Peninsular Malay, PAK – Pakistan, XIN – Xinjiang)

### 3.13 Haplogroup M24'41

Haplogroup M24'41 is subdivided into M24 and M41. M41 appears to be confined to south and eastern India (Thangaraj *et al.*, 2006; Chandrasekar *et al.*, 2009), dating to ~40 ka (using the coding-region mutation rate produced by Mishmar *et al.*, 2003) in Chandrasekar *et al.* (2009), although this mutation rate generally gives much older dates than the corrected mutation rate by Soares *et al.* (2009). Haplogroup **M24** is a very rare Sunda haplogroup which is invisible with HVS-I data but appears at present to be restricted to Southeast Asia, with a basal lineage in Vietnam and a subclade found in both Vietnam and north-western ISEA (Figure 3.40). It was first reported and named by Scholes *et al.* (2011) on Palawan, the Philippines. M24 dates to the start of LGM ~25 ka. Its subclade dates to ~12.5 ka, and is shared by lineages from Vietnam, North Borneo of Malaysia (Archaeogenetics Research Group, Huddersfield) and the Philippines (Scholes *et al.*, 2011). The Philippine lineage is sampled from one of the three non-negrito indigenous tribes called Tagbanua on Palawan, who practise small-scale agriculture and exhibit recent admixture with the Batak (Migliano *et al.*, 2007). According to Scholes *et al.* (2011), haplogroup M24 in the Philippines is restricted

to the Tagbanua only. The whole-mtDNA tree shows M24 is present in SEA at the LGM, ~25 ka and a late Pleistocene dispersal ~12.5 ka in ISEA.

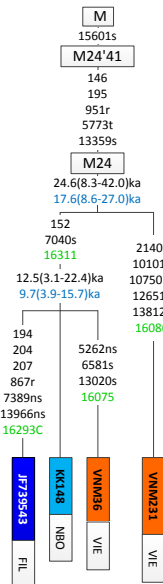


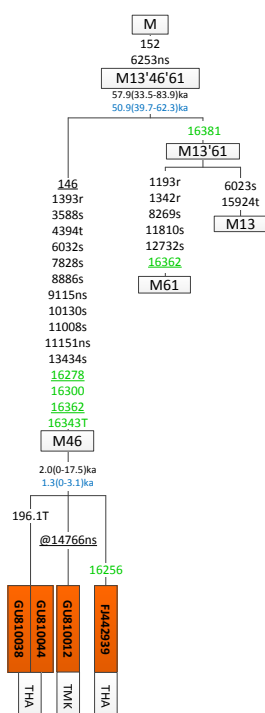
Figure 3.40 The tree of haplogroup M24'41. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, NBO – North Borneo, VIE – Vietnam)

### 3.14 Haplogroup M13'46'61

Haplogroup **M13'46'61** dates to ~58 ka, with a possible origin in Southeast Asia/South China, where there is most of the deepest diversity in the tree, and the subclades spread from here in different directions, in particular, northeast into northern China and Japan as M13a1, India and Tibet as M61a and M13b2, and lastly, M13b1 in Peninsular and Aboriginal Malays and ISEA. The whole-mtDNA tree includes 30 complete sequences: 4 M46, 8 M61 and 18 M13. (Note that I have amended certain branches of the tree but due to time constraints, dates are estimated mainly by  $\rho$  although certain ML dates are kept on the tree.)

M13'46'61 is subdivided into two basal subclades consisting of M46 and M13'61. **M46** (Figure 3.41), despite diverging ~50 ka, dates to only ~2 ka and is very rare; it is restricted to Thailand and the Moken (Dancause *et al.*, 2009; Pradutkanchana, Ishida and Kimura, 2010). The Moken, also known as the sea gypsies of the Andaman Sea, speak a language belonging to the Malayo-Polynesian branch of the Austronesian language family (Larish, 1999; Gordon, 2005). It is therefore of interest that their mtDNA affiliation (which also includes a preponderance of M21d lineages) relates more to their geography than their language. They

live off the coast of Myanmar and Thailand, subsisting through maritime foraging (Dancouse *et al.*, 2009). M46 has clearly undergone heavy drift, perhaps due to population subdivision in Thailand.

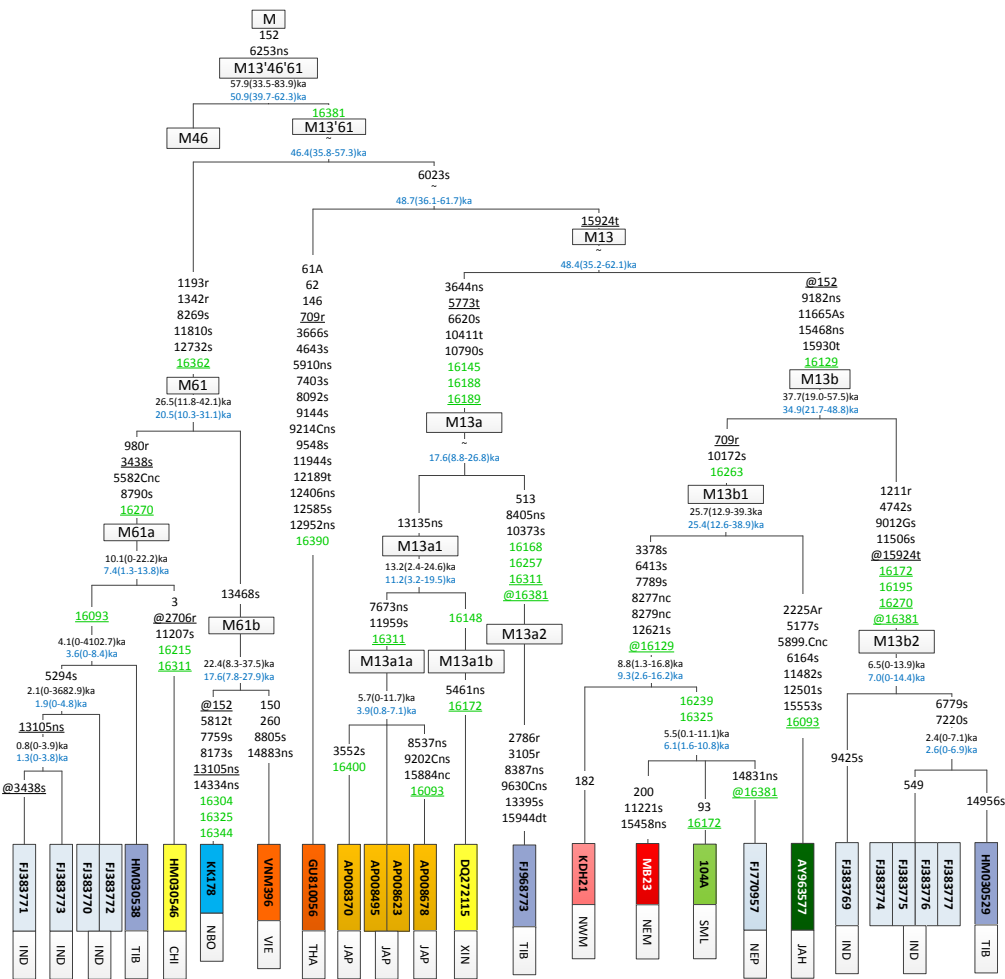


**Figure 3.41** The tree of haplogroup M46 nested within M13'46'61. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (THA – Thailand, TMK – Thailand Moken)

Haplogroup M13 shares a transition at np 16381 with M61, and the putative node (**M13'61**) dates to ~46 ka by  $\rho$ , and >50 ka by ML (Figure 3.41). The control-region position 16381 is a slow site, occurring five times compared to np 16362, 67 times, in a worldwide mtDNA tree (Soares *et al.*, 2009). Therefore, contrary to PhyloTree Build 15 (2012), M61 more likely clusters with M13 via np 16381 than with M46 via np 16362. There are three reversions of np 16381 seen in the Nepalese and Tibetan lineages within subclades M13a2 and M13b, suggesting that position 16381 may revert faster than it mutates forwards.

**M61**, dates to  $\sim 27$  ka ( $\rho \sim 21$  ka), and its subclades M61a and M61b each has different region distribution. **M61a** has a coalescence age  $\sim 10$  ka ( $\rho \sim 7$  ka) and is found in Yunnan, South China (Kong *et al.*, 2011), apparently with a subsequent Late Holocene spread events into Tibet (Kong *et al.*, 2011) and India (Chandrasekar *et al.*, 2009) by  $\sim 4$  ka. In HVS-I, M61 and M61a appear to be centred on South China. **M61b** is seen in only two individuals, from Vietnam and North Borneo (Archaeogenetics Research Group, Huddersfield), diverging  $\sim 22$  ka ( $\rho \sim 18$  ka); in the HVS-I database, M61b1 is seen only in Vietnam.

A **pre-M13** node defined by a transition at np 6023, dating to ~49 ka ( $\rho$ ) (~52 ka with ML), is found in a single individual from Thailand (Pradutkanchana, Ishida and Kimura, 2010), providing weak support for an origin in MSEA. **M13** dates to ~48 ka ( $\rho$ ) (~52 ka with ML) and it splits into M13a and M13b (Figure 3.42). **M13a** dates to ~18 ka ( $\rho$ ) and is found only in East Asia; the main subclade **M13a1** dates to the late Pleistocene ~13 ka (~11 ka by  $\rho$ ) and seems to be restricted to north China (Kong *et al.*, 2006) and Japan (Tanaka *et al.*, 2004); the latter forms a subclade, M13a1a, that dates to ~6 ka ( $\rho$  ~4 ka). Subclade M13a2 is found in a single individual from Tibet (Zhao *et al.*, 2009).



**Figure 3.42** The tree of haplogroups M46 and M13'61. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. Certain branches of the tree are changed but some of the ML dates are kept on the tree. (CHI – China, IND – India, JAH – Semang Jahai, JAP – Japan, NEM – Northeast Peninsular Malay, NBO – North Borneo, NEP – Nepal, NWM – Northwest Peninsular Malay, SML – Aboriginal Malay Semelai, THA – Thailand, TIB – Tibet, VIE – Vietnam, XIN – Xinjiang, China)

**M13b** dates to ~38 ka and is divided into M13b1 (~26 ka) and M13b2 (~7 ka) (Figure 3.42). A subclade nested within M13b1, dating to ~9 ka, is found in the Northwest Peninsular Malay, and further nested within a subclade (~6 ka) consists of lineages from Northeast

Peninsular Malaysia, Aboriginal Malay Semelai (this study) and Nepal (Fornarino *et al.*, 2009). **M13b2**, dating to ~6.5 ka, is restricted to India (Chandrasekar *et al.*, 2009) and then ~2.4 ka a subclade is further formed and shared between India and Tibet (Kong *et al.*, 2011).

The HVS-I network (see Figure 3.10 in Section 3.3) shows a basal lineage of M13b1 (previously M21b, but now recognisable as M13b1 in the HVS-I network) is reported in the Semang, Senoi (Hill, 2005) and Peninsular Malay, while the derivative types are found in Thailand, Sumatra, Borneo, Sulawesi, Aboriginal Malay, and Peninsular Malay (this study). This distribution in the HVS-I data is reflected in the whole-mtDNA analysis, where a basal lineage of M13b1 is seen in a Semang Jahai individual (Macaulay *et al.*, 2005). An unclassified branch of M13 has also been seen shared between a Peninsula Malay and a Thai individual.

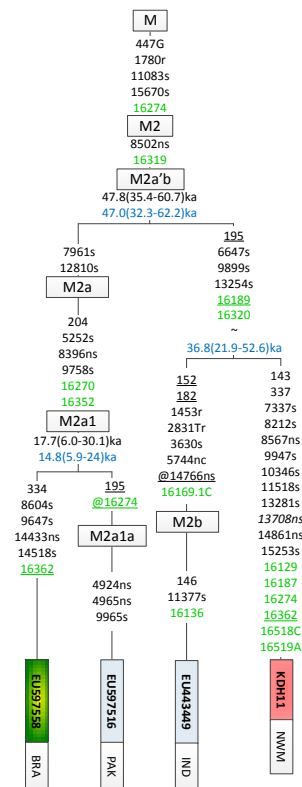
Considering the basal pre-M13 lineage and the oldest subclade, M13b1, dating to ~26 ka, are found in Thailand and the Malay Peninsula respectively, with clades in Japan and South Asia dating to the mid-Holocene, M13 may have had a MSEA/Sunda origin with several comparatively recent offshoots in different directions into North China/Japan as M13a1 by ~13 ka and M13b2 in India and Tibet by ~6.5 ka. In fact, the whole haplogroup M13'46'61 can be argued to be a deep Sunda haplogroup. Firstly, M61b (dates to the Late Glacial ~18-22 ka) is found in North Borneo and Vietnam, in what looks like the relict descendants of the first settlements on Sunda shelf. Secondly, M46 evolved earlier than M13'61 and is restricted to Thailand. In a similar pattern to subclades M13a1 and M13b2, M61a would have spreads northwards by ~10 ka into China and Tibet, to finally reach India by ~2 ka.

### 3.15 Haplogroup M2

Haplogroup **M2a'b** is basal to haplogroup M and dates to ~48 ka, reconstructed here by three complete sequences (Figure 3.43). M2 has been studied extensively by Kumar *et al.* (2008) and 76 complete sequences from India were used in their phylogenetic analysis. M2 is found widely distributed in India at high frequency (~10% of Indian M haplogroups; Bamshad *et al.*, 2001; Metspalu *et al.*, 2004), which is significantly more pronounced in southern India compared to north and possibly represent the earliest settlers in South Asia who colonised India through the southern coastal route. However, these complete sequences were not included in this study due to time constraints.



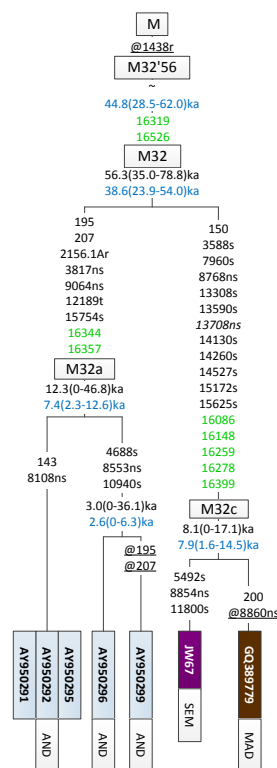
**M2a1** has a region-specific distribution in India, mainly among the Indo-European speakers in western and central India (Kumar *et al.*, 2008). In Figure 3.43, M2a1 dates to ~18 ka and they are from Pakistan and Brazil (Hartmann *et al.*, 2009). The presence of a sample from Brazil in the Indian haplogroup M2a, as well as R7a by Hartmann *et al.*, (2009), is intriguing. Since it is only a single instance, this sample could have possibly migrated recently from India into South America. **M2b** is found at high frequency in Dravidian speakers of central and south India, and in the Korku, an Austro-Asiatic tribe of central India, but is virtually absent among Indo-European speakers of western and central India as well as Tibeto-Burman speakers of north-east India (Kumar *et al.*, 2008). In this study, a Northwest Peninsular Malay shares six defining mutations with M2b. The pre-M2b Malay lineage represents the relict descendant in Peninsular Malaysia (although a Malay speaker) from the first settlers that arrived via the southern route from India ~37 ka (by  $\rho$ ). On the other hand, the Malay HVS-I sequence almost matches one in the Apatani tribe in eastern India, so the split is likely to be very recent.



**Figure 3.43** The tree of haplogroup M2. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BRA – Brazil, NWM – Northwest Peninsular Malay, PAK – Pakistan)

### 3.16 Haplogroup M32'56

Haplogroup M32'56 subdivides into M32 and M56. M56 is reported in the Austro-Asiatic-speaking Korku tribe in central India, dating to ~15 ka using the Mishmar *et al.* coding-region mutation rate (Chandrasekar *et al.*, 2009). **M32** is a rare haplogroup, dating to ~56 ka, and is subdivided into M32a and M32c (Figure 3.44). **M32a** appears to be exclusively restricted to Andaman Islands (Thangaraj *et al.*, 2006) that have undergone high genetic drift, possibly resulting in young dates of ~12 ka, with a subclade nested within that dates to ~3 ka. **M32c** also experienced high drift, dating to ~8 ka and it is shared between a Southeast Peninsular Malay (this study) and interestingly in Madagascar (Dubut *et al.*, 2009).



**Figure 3.44** The tree of haplogroup M32. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AND – Andaman Islands, MAD – Madagascar, SEM – Southeast Peninsular Malay)

The native Madagascar speak Austronesian-languages, which belonged to the most widespread language family in the world, with a distribution extending more than half way around the globe from Madagascar to Easter Island (Bellwood *et al.*, 2006). The Malagasy language in Madagascar nested within the Barito subgroup, on South Borneo, which was believed to have been in Madagascar since 400-500 A.D., although Adelaar (1994) claimed that this date is at least two centuries too early. Although it is impossible to know a direction

of migration with minimal complete sequences, the presence of M32c in Madagascar pre-dates the Austronesian-speakers expansion to mid-Holocene.

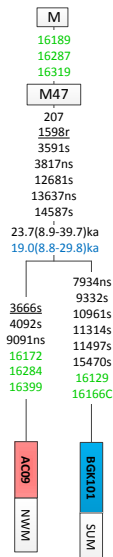


Figure 3.45 The tree of haplogroup M47. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka.

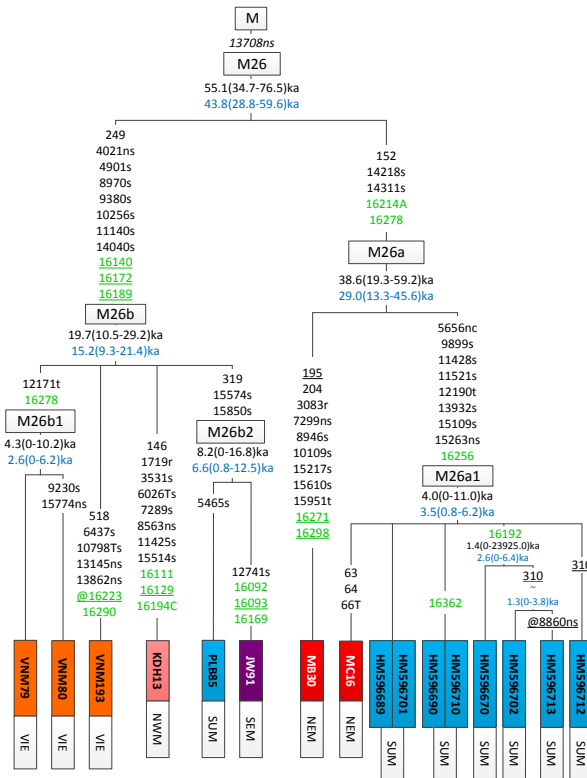
### 3.17 Haplogroup M47

M47 is a rare, possibly Sunda haplogroup in SEA (Figure 3.45). It dates to ~24 ka and found in Bangka, Sumatra (Archaeogenetics Research Group, Huddersfield) and Northwest Peninsular Malaysia (this study). There is potentially another M47 sample from Pekanbaru, Sumatra with similar HVS-I motif to the Bangka lineage and not seen anywhere else by the HVS-I database. Due to the low number of whole mtDNA sequences, the direction of migration remains unclear.

### 3.18 Haplogroup M26

**M26** is a rare clade that dates to ~55 ka and divides into M26a and M26b (Figure 3.46), both of which have distinctive HVS-I motifs. M26 has a deep and widespread Sunda distribution, across MSEA/Malay Peninsula and Sumatra. **M26a** dates to ~39 ka, with a basal lineage in a Northeast Peninsular Malay (this study), and a derived subclade, **M26a1**, also seen in Sumatra (Gunnarsdóttir *et al.*, 2011b), dates to ~4 ka. In the HVS-I database, both

derived and underived lineages are also seen (although very rarely) in Thailand and Eastern Indonesia.



**Figure 3.46** The tree of haplogroup M26. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, VIE – Vietnam)

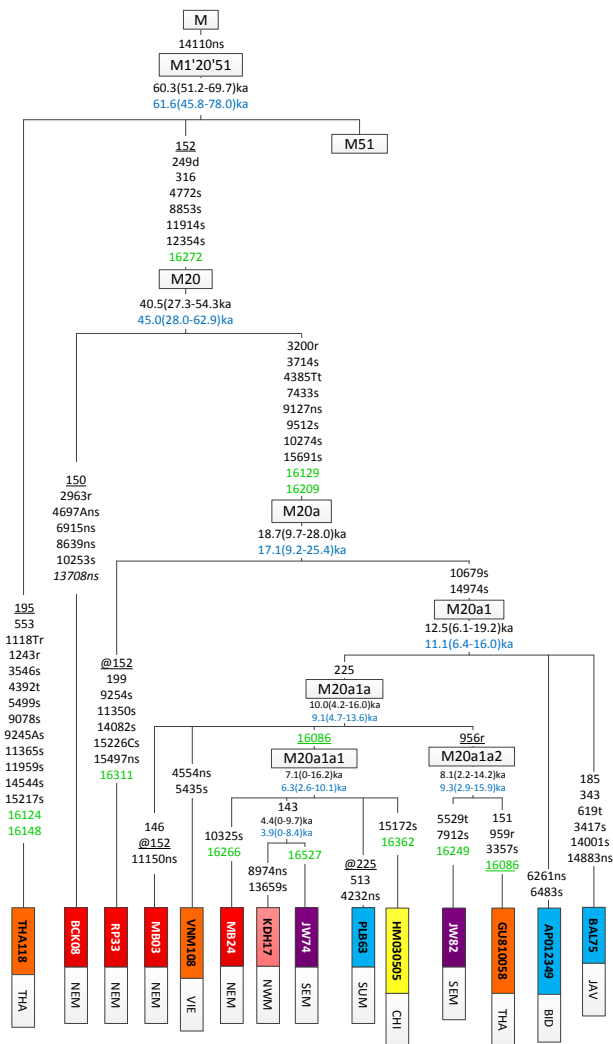
**M26b** dates to ~20 ka, with basal lineages in Vietnam and Northwest Peninsular Malaysia. There are two subclades, M26b1 and M26b2. **M26b1** dates to ~4 ka and is seen in Vietnam only in the whole-mtDNA tree (Archaeogenetics Research Group, Huddersfield), but with single examples also in Sumatra and Lombok in the HVS-I database. **M26b2** has an older date of ~8 ka and is found in Southeast Peninsular Malaysia (this study) and Sumatra (Archaeogenetics Research Group, Huddersfield). The distribution and age of M26 suggests an ancient root in the Sunda region dating to the first settlers ~55 ka. Although sample sizes are presently small, M26a1 could be a potential marker of Bellwood's hypothesis (1990) regarding the Neolithic expansion of farmers along the coastlines dispersal from Peninsular Malaysia into Sumatra; although Bellwood no longer believes this (personal communication).

### 3.19 Haplogroup M1'20'51

The putative haplogroup M1'20'51 dates to ~60 ka, defined by a transition at np 14110 (which occurs only three times in a global phylogeny: Soares *et al.*, 2009) and shared by haplogroups M1, M20 and M51. Relevant to this study are subclades M20 and M51 shown in Figure 3.47. There are 40 complete sequences on the tree: 13 belong to M20 and 26 to M51, and there is basal paraphyletic M1'20'51 lineage in Thailand (seen twice in Thailand in the HVS-I database). All subclade names are assigned here for the first time apart from subclades M51a to M51a1a (Peng *et al.*, 2010). M20 and M51 both have deep roots within the Sunda regions, especially MSEA/Peninsular Malaysia. This deep ancestry may also point to an ultimately Southeast Asian source for the enigmatic Mediterranean and East African haplogroup M1 (Olivieri *et al.*, 2006).

**M20** dates to ~41 ka and **M20a** to the end of LGM ~19 ka. Both have basal lineages seen in Northeast Peninsular Malay, indicating an origin lies somewhere in MSEA, and M20a is mainly seen in MSEA/South China in the HVS-I database, with very few in ISEA. **M20a1** dating to before the second flooding of the Sunda shelf ~12.5 ka, includes basal lineages in the Bidayuh of Sarawak in North Borneo (Jinam *et al.*, 2012) and Java (Archaeogenetics Research Group, Huddersfield). **M20a1a**, dating to ~10 ka, is divided into M20a1a1 and M20a1a2, with the basal lineages in Northeast Peninsular Malaysia (this study) and Vietnam (Archaeogenetics Research Group, Huddersfield). **M20a1a1** dates to ~7 ka, where it is found in South China (Kong *et al.*, 2011), Peninsular Malay (this study) and Sumatra (Archaeogenetics Research Group, Huddersfield). **M20a1a2**, dates to ~8 ka, is seen in Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Southeast Peninsular Malaysia (this study).

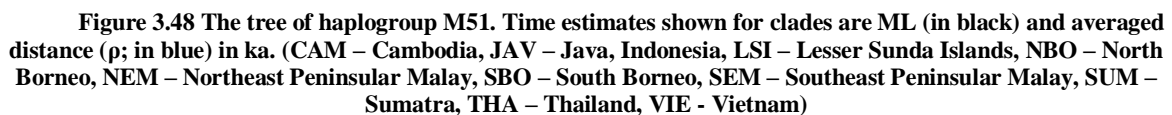
Subclade M20a1a2 is found in single individuals from Thailand and Peninsular Malaysia, while M20a1a1 is seen in several Peninsular Malay, and one each from Sumatra and South China. However, the age difference is not significant given the confidence intervals of estimation. The pattern suggests an ancient source centred on MSEA and most probably recent dispersals northwards into South China, and southwards in Sumatra, perhaps relating to the flooding of the Strait of Malacca in the early Holocene (Oppenheimer, 1998).



**Figure 3.47** The tree of haplogroup M20. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BID – Sarawak Bidayuh, CHI – China, JAV – Java, Indonesia, NBO – North Borneo, NEM – Northeast Peninsular Malay, Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, THA – Thailand, VIE – Vietnam)

**M51** dates to ~37 ka and divides into M51a and M51b (Figure 3.48), which are seen across the Sunda area. **M51a** dates to ~33 ka, which again further subdivides into M51a1 (~25.5 ka) and M51a2 (~31 ka). **M51a1** dates to the LGM, its subclade **M51a1a** (~10 ka) is found in Vietnam (Peng *et al.*, 2010) and a recent subclade nested within, dating to ~1 ka, is seen localised in Sumatra, suggesting a very recent dispersal from MSEA (Gunnarsdóttir *et al.*, 2011b). **M51a1b** dates to ~10 ka and is seen in Java, Thailand (Archaeogenetics Research Group, Huddersfield) and Cambodia (Hartmann *et al.*, 2009).

The older subclade **M51a2** pre-dates the LGM and has a basal lineage in Northeast Peninsular Malaysia (this study) and a subclade, **M51a2a**, dating to ~17 ka, shared between an instance from South Borneo (Archaeogenetics Research Group, Huddersfield) and two from Vietnam (Peng *et al.*, 2010), the latter formed a further subclade, **M51a2a1** (~7 ka).



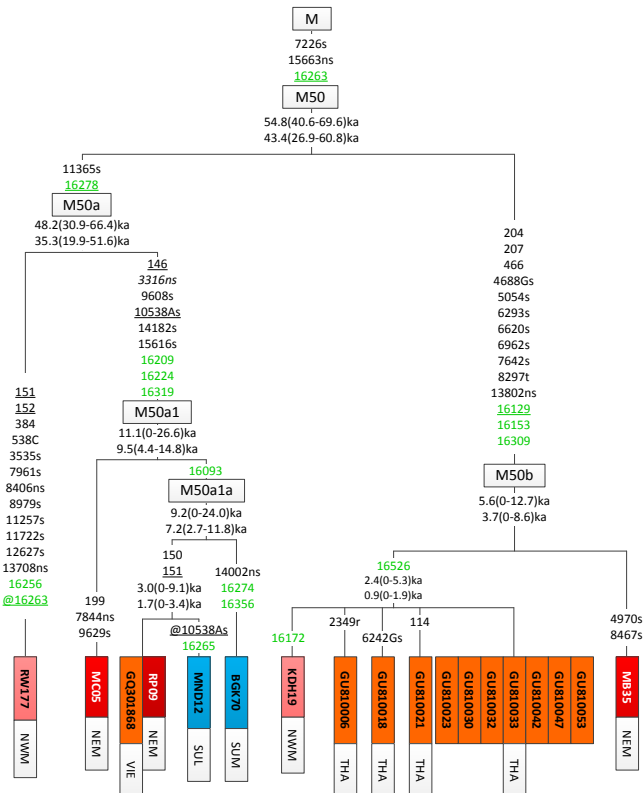
Basal lineages of both **M51b2** and **M51b2a** are seen in Northeast Peninsular Malay (this study), separating towards the end of the LGM, ~19 ka (~20 ka and ~18 ka by  $\rho$  respectively). **M51b2a1** dates to ~14 ka with a basal lineage in Mataram (in the Lesser Sunda Islands), and a subclade, M51b2a1a, dating to ~2 ka and seen in Vietnam (Archaeogenetics Research Group, Huddersfield).

We can confirm the overall distribution with HVS-I data. Peng *et al.* (2010) found several unclassified haplogroups observed in their Cham and the Kinh Vietnamese samples to cluster with previously reported (more than 3000) unclassified mtDNA HVS-I sequences from ISEA, and one of the haplogroups they recognised was M51. According to their reduced-median HVS-I network (Fig. 3 in Peng *et al.*, 2010) M51 is widely distributed almost exclusively in Southeast Asia, in particular the southern pole of MSEA such as Vietnam, Thailand, Peninsular Malaysia, and Borneo and Indonesia in ISEA. Haplogroup M51a1, recognisable by a transition at np 16294, has root types found in Indonesia, Cambodia and Thailand. Its derivatives are restricted to MSEA in Cambodia, Thailand and Vietnam. Haplogroup M51b1, defined by a transition at np 16311, has the root type seen in Vietnam and Indonesia, and the derived types are found in Vietnam, Peninsular Malaysia and Indonesia. Lastly, haplogroup M51b2 is also recognisable on the network, defined by a transition at np 16189. M51b2 is predominantly found at high diversity in Indonesia, with minor sharing of the root type with Vietnam (the exact breakdown of frequency was not available in Peng *et al.*, 2010). Its derivatives are also found in Peninsular Malaysia. The HVS-I signals are therefore similar to those observed in the whole-mtDNA tree: M51, similar to M20, is a deep Sunda haplogroup in MSEA and widely distributed throughout ISEA spread from the beginning of the Late Glacial onwards, the time at which sea-levels began to rise – possibly implying an impact of sea-level rise in MSEA comparable to that in ISEA.

### 3.20 Haplogroup M50

**M50** spread through the Sunda region, dating to ~55 ka and split into M50a and M50b, both of which appear to have their origins in MSEA (Figure 3.49). **M50a** dates to ~48 ka, and it is seen in Northwest Peninsular Malay (this study). The lineage appears also in Northeast Peninsular Malaysia as subclade **M50a1** ~11 ka, from where it may have spread into Vietnam and Sumatra, Indonesia (Archaeogenetics Research Group, Huddersfield) as **M50a1a** ~9 ka, a further subclade of M50a1a, aged around 3 ka, indicates lineages in Vietnam (Peng *et al.*, 2010) and off the southeastern Sunda shelf in Sulawesi, Indonesia (Archaeogenetics Research Group, Huddersfield).





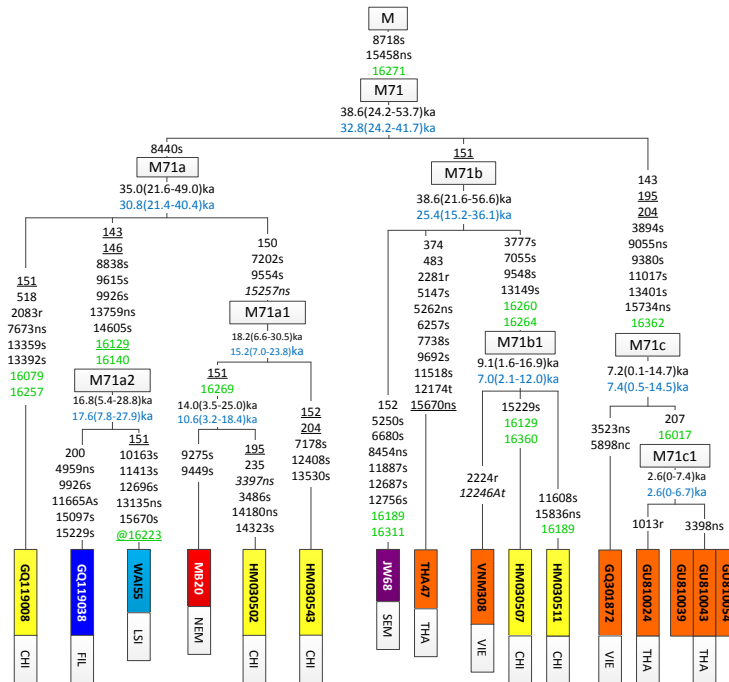
**Figure 3.49** The tree of haplogroup M50. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, VIE – Vietnam)

The long branch of **M50b** indicates high drift resulting in a date of ~6 ka, and is also characteristic to Northeast and Northwest Peninsular Malaysia (this study). A further subclade defined by a transition at np 16526, dating to ~2 ka, and is found mainly in Thailand (Pradutkanchana, Ishida and Kimura, 2010) but also Northwest Peninsular Malay (this study), suggesting recent northern spread.

### 3.21 Haplogroup M71

Haplogroup **M71** dates to ~39 ka and is subdivided into three subclades, M71a, M71b and M71c. M71 appears to have an early origin in northern mainland SEA/South China with subsequent spread on the Sunda shelf. The phylogeny of M71 includes 16 complete sequences (Figure 3.50). **M71a**, dating to ~35 ka, is widely distributed in South China, the Sunda shelf and beyond, with the basal lineage seen in South China (Tabbada *et al.*, 2010). It is further divided into M71a1 and M71a2. M71a1, dating to ~18 ka, is seen in South China. A subclade nested within M71a1 is found in South China (Kong *et al.*, 2011) and Northeast Peninsular Malay (this study). **M71a2** dates to ~17 ka, and is restricted to ISEA in the

Philippines (Tabbada *et al.*, 2010) and Sumba of Lesser Sunda Islands (Archaeogenetics Research Group, Huddersfield).



**Figure 3.50** The tree of haplogroup M71. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, SEM – Southeast Peninsular Malay, THA – Thailand, VIE – Vietnam)

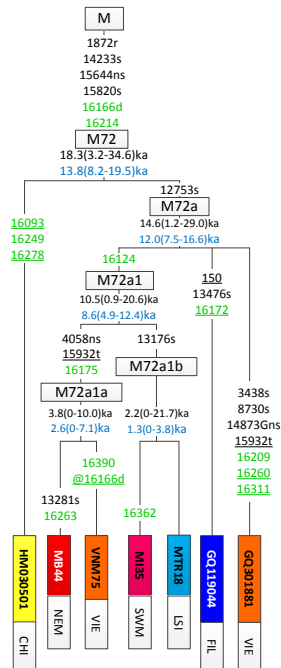
**M71b** is so far confined to MSEA and South China, dating to ~39 ka, with the basal lineages being found in Thailand (Archaeogenetics Research Group, Huddersfield) and Southeast Peninsular Malay (this study). A subclade **M71b1** dates to ~9 ka, and appears in Vietnam (Archaeogenetics Research Group, Huddersfield) and South China (Kong *et al.*, 2011).

The third M71 branch, **M71c** dates to ~7 ka, is so far confined to northern MSEA in Vietnam as a basal branch (Peng *et al.*, 2010) and in Thailand (Pradutkanchana, Ishida and Kimura, 2010) as **M71c1** ~3 ka.

### 3.22 Haplogroup M72

Haplogroup **M72** is another primary M branch based in northern MSEA/South China (Figure 3.51). It experienced genetic drift with its expansion dating to the LGM (~18 ka), and Sunda spread from the Late Glacial onwards. A basal lineage of M72 is seen in South China (Kong *et al.*, 2011). This haplogroup is seen, albeit at low levels, throughout the Sunda shelf and beyond as clustered subclades. **M72a** dates to ~15 ka with basal lineages found both in

Vietnam (Tabbada *et al.*, 2010) and the Philippines (Peng *et al.*, 2010). **M72a1**, dating to ~11 ka, is divided into **M72a1a** ~4 ka, and **M72a1b** ~2 ka. M72a1a shows local MSEA spread in Vietnam (Archaeogenetics Research Group, Huddersfield) and Northeast Peninsular Malay (this study), while M72a1b is detected in Southwest Peninsular Malay (this study) and Mataram, Indonesia (Archaeogenetics Research Group, Huddersfield).



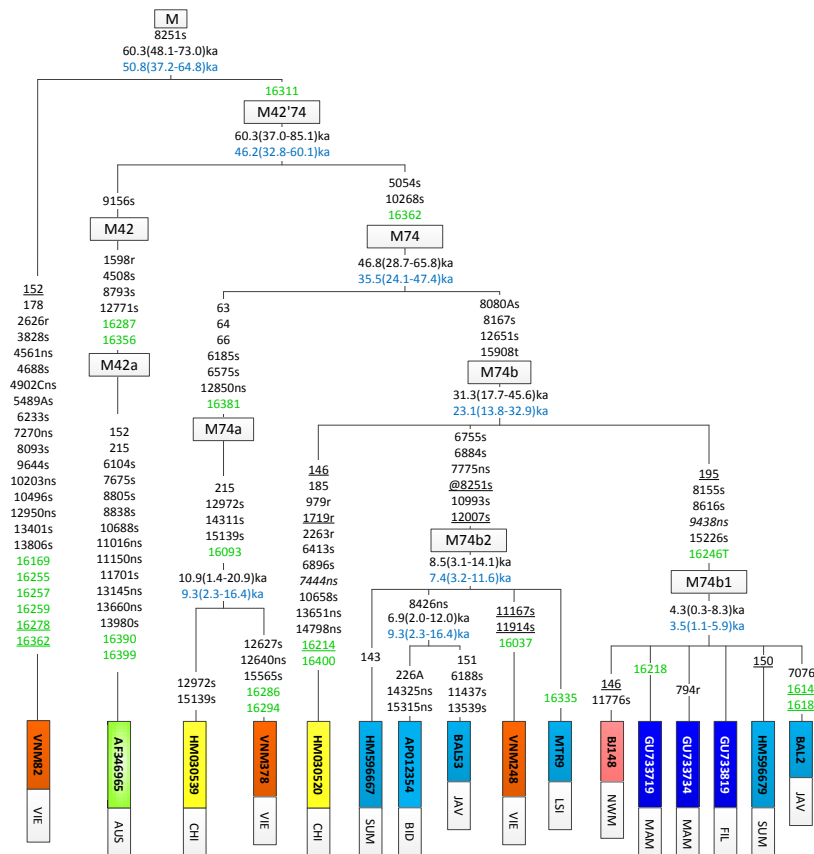
**Figure 3.51** The tree of haplogroup M72. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, NEM - Northeast Peninsular Malay, SWM - Southwest Peninsular Malay, VIE - Vietnam)

### 3.23 Haplogroup M42'74

Haplogroup **M42'74** dates to ~60 ka, and appears to be of early Sunda origin with very early spread throughout ISEA, on southwards to Australia and later north to South China. It is divided into M42 and M74 (Figure 3.52). A complete sequence from Vietnam (Archaeogenetics Research Group, Huddersfield) shares a transition at np 8251 which defines the pre-M42'74 node, and dates to ~60 ka too. M42 (or **M42a**) is represented here by a singleton from an Australian Aborigine (Ingman *et al.*, 2000). The link may however be spurious, as np 8251, which is the sole position defining the combined clade, is very fast-evolving (Soares *et al.*, 2009). Haplogroup M74 includes 14 complete sequences, and is seen commonly from China through MSEA and ISEA to the Philippines.

**M74** dates to ~47 ka, and has two subclades: M74a and M74b. **M74a** dates to ~11 ka, and is, so far, only seen in South China (Kong *et al.*, 2011) and Vietnam (Archaeogenetics Research Group, Huddersfield). **M74b** dates to ~31 ka and while a basal lineage is seen in Yunnan China (Kong *et al.*, 2011), two nested subclades are confined to the Sunda shelf. One of these, **M74b1** dates to ~4 ka and is seen both in Northwest Peninsular Malay (this study) and different locations of ISEA: Sumatra (Gunnarsdóttir *et al.*, 2011b) and Java, Indonesia (Archaeogenetics Research Group, Huddersfield), and the Mamanwa and Philippines (Gunnarsdóttir *et al.*, 2011a).

**M74b2** dates to ~9 ka, with basal branches in Vietnam, as well as Mataram and Sumatra in Indonesia (Gunnarsdóttir *et al.*, 2011b; Archaeogenetics Research Group, Huddersfield). A further subclade nested within, defined by a transition at np 8426, dating to ~7 ka and is seen in ISEA: Java, Indonesia (Archaeogenetics Research Group, Huddersfield) and a Bidayuh in Sarawak in North Borneo (Jinam *et al.*, 2012).



**Figure 3.52** The tree of haplogroup M42\*74. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AUS – Australia, BID – Sarawak Bidayuh, CHI – China, FIL – Philippines, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, SUM – Sumatra, VIE – Vietnam)

### 3.24 Haplogroup M73'79

Haplogroup **M73'79** is a deep primary Sunda lineage, spreading widely and early in SEA. It dates to ~61 ka, and divides into M73 (~45 ka) and M79 (~21 ka). The phylogenetic tree of M73'79 includes 14 complete sequences: eleven M73 and three M79 (Figure 3.53).

**M73a** dates to ~30 ka, and subdivides into M73a1 and M73a2, the former in MSEA and the latter in ISEA. **M73a1**, dating to ~20 ka, and is seen in Vietnam (Tabbada *et al.*, 2010) and Thailand (Peng *et al.*, 2010). **M73a2**, dating to ~26 ka, and is found in Alor, Indonesia (Archaeogenetics Research Group, Huddersfield) and the Philippines (Tabbada *et al.*, 2010).

**M73b**, dating to ~39 ka, is seen mainly in ISEA: in North Borneo and Sumatra, Indonesia (Archaeogenetics Research Group, Huddersfield) as well as in MSEA. **M73b1** dates to ~29 ka and a basal branch is again seen in MSEA: Vietnam (Archaeogenetics Research Group, Huddersfield). A further subclade, **M73b1a**, dating to ~11 ka, and is further divided into two branches, one in ISEA, the other in MSEA. The first branch dates to ~5 ka and is found in Brunei, North Borneo (Archaeogenetics Research Group, Huddersfield), Sumatra and elsewhere in Indonesia (Tabbada *et al.*, 2010), while the second branch, ~11 ka, is seen in Vietnam (Peng *et al.*, 2010) and a Southwest Peninsular Malay (this study).

We now look at the HVS-I network of haplogroup M73 (Figure 3 in Peng *et al.*, 2010). Haplogroup M73a, recognisable by a transversion at np 16184A in the figure, has its root types found in Peninsular Malaysia and Indonesia, and the derivatives are seen in Cambodia and Thailand. Haplogroup M73b1a, recognisable by a transition at np 16354, is found in Vietnam and Peninsular Malaysia. The results corresponded with the whole-mtDNA tree that M73 indeed have origins confined to SEA.

M79 is a rare haplogroup that dates to the LGM ~21 ka, where a basal lineage is seen in Yunnan, South China (Kong *et al.*, 2011). A subclade nested within, dating to ~20 ka, is shared between Vietnam and Java, Indonesia (Archaeogenetics Research Group, Huddersfield). Similar to M73, M79 is restricted to the relict descendant on the Sunda shelf since the LGM.

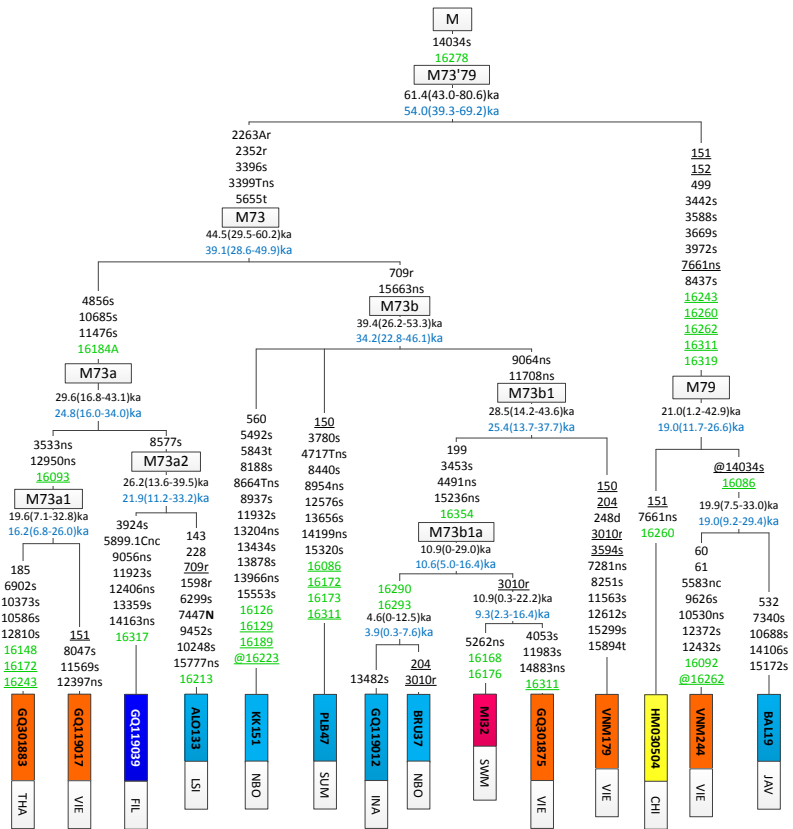


Figure 3.53 The tree of haplogroup M73'79. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, INA – Indonesia, JAV – Java, Indonesia, LSI – Lesser Sunda Islands, NBO – North Borneo, SUM – Sumatra, THA – Thailand, VIE – Vietnam)

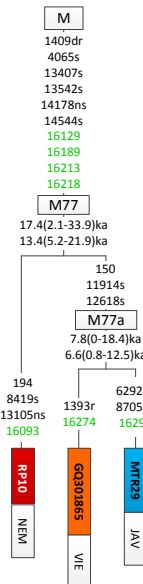


Figure 3.54 The tree of haplogroup M77. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (JAV – Java, Indonesia, NEM – Northeast Peninsular Malay, VIE – Vietnam)

### 3.25 Haplogroup M77

Haplogroup **M77** is a rare primary Sunda M haplogroup that has undergone genetic drift and found both in MSEA and ISEA (Figure 3.54). The HVS-I sequences for M77 are reported so far in Vietnam and Peninsular Malaysia only (Peng *et al.*, 2010). Meanwhile, my results show that this haplogroup dates to ~17 ka, and a basal lineage is found in a Northeast Peninsular Malay (this study). A Holocene subclade **M77a**, dating to ~8 ka, and is found both in MSEA (Vietnam: Peng *et al.*, 2010) and ISEA (Java, Indonesia: Archaeogenetics Research Group, Huddersfield).

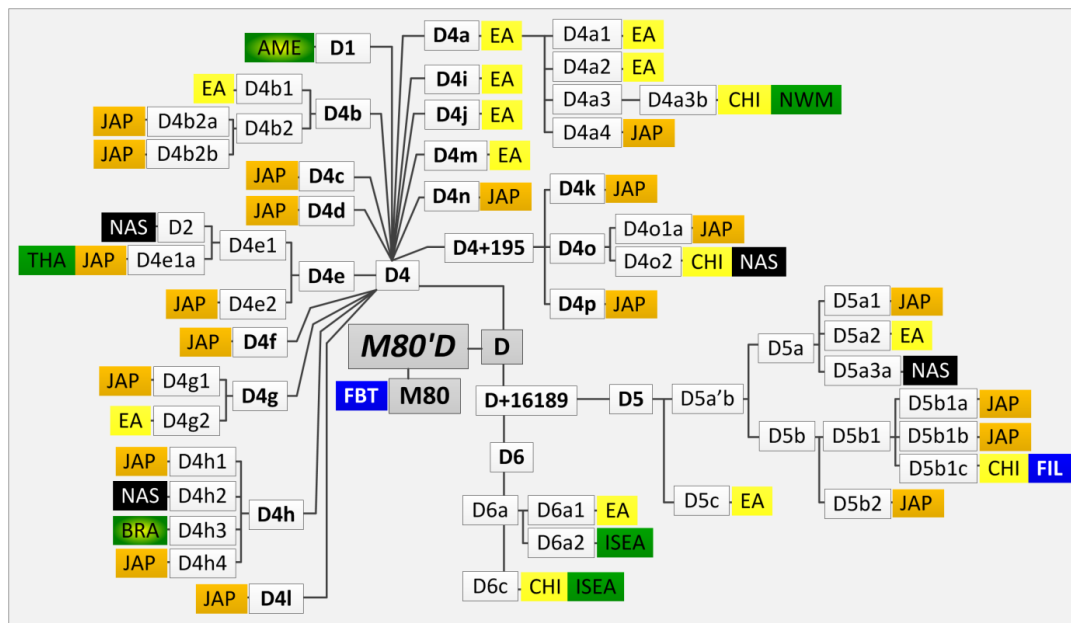
### 3.26 Haplogroup D

Haplogroup **M80'D** has been inferred (Scholes *et al.*, 2011; Phylotree by van Oven and Kayser, 2009) by a shared transition at np 4883 between a singleton M80 found among the Batak negrito of the Philippines and the very large D haplogroup (Figure 3.55). The latter is present in significant frequency and diversity in MSEA (based on the HVS-I data in Laos reported in Bodner *et al.*, 2011) and lower frequency throughout ISEA, while it is widespread and very common throughout East Asia and the New World (Derenko *et al.*, 2010). This background, and the single deep link with ISEA, raises the possibility that the ancestry of haplogroup D may originally have lain on the Sunda shelf, even the southern part of it.

Haplogroup D dates to ~52 ka, and has two major subclades: the primary branch D4 (includes D1) estimated at ~29 ka and D5'6, or D4+16189C which, if a true ancestral node, would date similarly to D overall, ~52 ka. D5 dates to ~43 ka and D6 ~31 ka. The phylogenetic tree of haplogroup D includes 339 complete sequences: 279 D4 (and D1), 50 D5 and 10 D6. I have also included 35 Chinese complete sequences (Zheng *et al.*, 2011) in the tree but they were not included in ML estimations because of time constraints.

Haplogroup D is mainly a Northern Asian haplogroup, encompasses almost 20% of the total mtDNA variation in most of northern Asia and retains a very high overall frequency in all regional northern Asian groups (11-34%), central Asian (14-20%) and eastern Asian (10-43%) populations (Derenko *et al.*, 2010), so that Northern Asian lineages occur throughout the tree. While few of them are specific for northern Asian populations, D is also very common in eastern, central Asia and America (ref e.g. Derenko *et al.*, 2010). D1, a subclade subsumed by D4, is found in the Americas. D4 is the most represented of D subclades and are

found at high frequencies in China, Japan and Korea (Mormina, 2007). D4e appears to have a root in Japan, where subclade D2 is found in North Asia, and D4e1a is found in Japan as well as Thailand. D4h has four subclades, D4h1 is almost entirely found in Japan and at much lower levels in China. D4h2, D4h3 and D4h4 are each represented by one complete sequence, and they are found respectively in North Asia, Brazil and Japan. D4a, D4i, D4j and D4m are widely found in China and Japan. Subclade D4a3b is found in China and Northwest Peninsular Malaysia. D4+195 includes subclades D4k, D4o and D4p, all showing root types in Japan, with the exception of D4o2 which is found in China and North Asia.



**Figure 3.55 Schematic diagram of haplogroup D and its major subclades distribution. (AME – America, BRA – Brazil, CHI – China, EA – East Asia, FBT – Philippines Batak, FIL – Philippines, ISEA – Island Southeast Asia, JAP – Japan, NAS – North Asia, NWM – Northwest Peninsular Malaysia, THA - Thailand)**

Haplogroup D5 and its subclades are also found mainly in China and Japan (Figure 3.55). However, D5a3a is seen rarely in North Asia, and D5b1c in China and the Philippines. It reaches the highest frequencies in Korea and Northeast China (Mormina, 2007). Haplogroup D6 is a small haplogroup and is divided into D6a and D6c. D6a1 is seen in China and Japan, D6a2 on the other hand is seen in ISEA. D6c is found in South China and then it spreads into ISEA. Additionally, the Laotian HVS-I data indeed complement the whole-mtDNA picture by showing the presence of D4b1b, D4b2b, D4e1'3, D4g2a, D5b, D5a2a1 and D\* in MSEA (Laos in Bodner *et al.*, 2011), when the whole-mtDNA trees appear to be mainly restricted to East Asia.



### 3.26.1 Haplogroup D4

In general, haplogroup D4a seems to have originated in China dating to the Late Glacial based on the paraphyletic basal lineages. The whole-mtDNA tree suggests several dispersals, some may be earlier and some later, from China to Japan roughly in the early to mid Holocene. In Figure 3.56, **D4a** is mainly seen in Japan (Tanaka *et al.*, 2004) and it dates to ~17 ka, but the basal lineages are restricted to China: one in the South and three in the North (Zheng *et al.*, 2011). D4a includes four subclades, D4a1, D4a2, D4a3 and D4a+C16294T (D4a4). **D4a1** dates to ~7 ka, and so far has been identified five subclades: D4a1a, D4a1b, D4a1c, D4a1d and D4a1e. D4a1e is represented here by one Japanese complete sequence. **D4a2** dates to ~11 ka, and subclade D4a2a around 2 ka. **D4a3** dates to ~14 ka, and subclades **D4a3a** remains seen in Japan (Tanaka *et al.*, 2004) ~3 ka. **D4a3b** is seen in northeastern China (Kong *et al.*, 2003a) and Northwest Peninsular Malaysia (this study) around 7 ka. **D4a+C16294T** dates to ~13 ka and is restricted to Japan. D4a3 types are also found in South China in the HVS database. The lineage of Northwest Peninsular Malay within D4a3b, similar to the Thai lineage within subclade D4e1a, can be seen as an intrusive dispersal from northern China in an upper bound of ~7 ka. However, seeing that they were extremely rare in Malaysia, these lineages could have arrived there quite recently. Detailed description for haplogroup D4 is available in Appendix E.

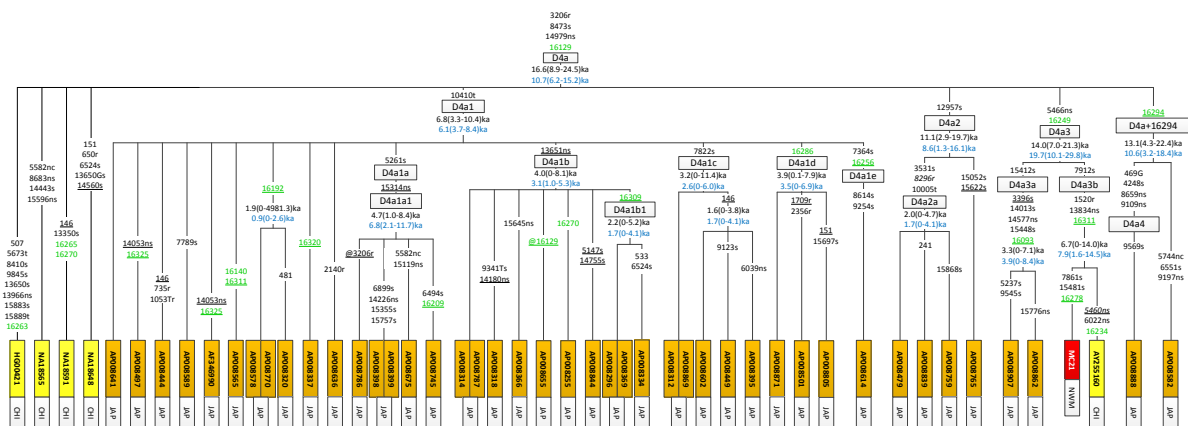


Figure 3.56 The tree of haplogroup D4a. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, JAP – Japan, NWM – Northwest Peninsular Malay)

### 3.26.2 Haplogroup D5

Haplogroups D5 (~43 ka) and D6 (~31 ka) share the ancestral node diverged from D and defined by a transition at np 16189 that dates (similar to haplogroup D overall) to ~52 ka. Haplogroup D5 is prevalent in China (<10%; Yao *et al.*, 2002a) and much lower frequency in

southern Siberia (1.5%; Derenko *et al.*, 2003), but rare in ISEA (~3 %) although it reaches >10% in some parts of Sulawesi (Hill *et al.*, 2007), or absent in Central Asia (Kolman *et al.*, 1996). D5a is found in Liaoning, Wuhan, Xinjiang and Qingdao in northern China (Yao *et al.*, 2002a). Detailed description is available in Appendix E.

**D5b** dates to ~27 ka and it is divided into subclades D5b1, D5b2 and newly named D5b3 (Figure 3.57). The root type of D5b is found in the Laotian HVS-I data (Bodner *et al.*, 2011). **D5b1** dates to ~20 ka and it includes D5b1a, D5b1b, D5b1c and D5b1d, all restricted to Japan and China. Both subclades **D5b1a** and **D5b1b**, dating to ~12 ka and ~13 ka respectively, are found only in Japan (Tanaka *et al.*, 2004). **D5b1c** has a basal lineage found in Yunnan, South China (Kong *et al.*, 2003a) and its subclade D5b1c is seen in north China (Zheng *et al.*, 2011) and the Philippines (Tabbada *et al.*, 2010). **D5b1d** is found in two instances from northern China (Zheng *et al.*, 2011). The HVS-I database shows that D5b1c (previously classified as D5d1 and dated to ~4 ka in Hill *et al.*, 2007) is also seen quite frequently in Indonesia, especially Sulawesi and Taiwan. The root type of this branch (D5b1c) is not found in Taiwan, but three derived types are found there, suggesting that the root type may have been lost due to drift. Although there are very few whole-mtDNAs in D5b1c, considering the HVS-I database and as suggested by Hill *et al.* (2007), it can be plausibly ascribed to a mid-Holocene Out of Taiwan event through the Philippines into ISEA.

**D5b2** dates to ~10 ka and found only in two individuals from Japan (Tanaka *et al.*, 2004). **D5b3** dates to ~5 ka and it is found in one individual each from Thailand (Archaeogenetics Research Group, Huddersfield) and Southeast Peninsular Malaysia (this study). D5b3 is not recognisable by the HVS-I motifs alone, and since there are only two instances within D5b3, it is difficult to phylogenetically infer any further. **D5c** is a small subclade found in China (Zheng *et al.*, 2011) and two individuals in Japan (Tanaka *et al.*, 2004) ~3 ka as **D5c1**.

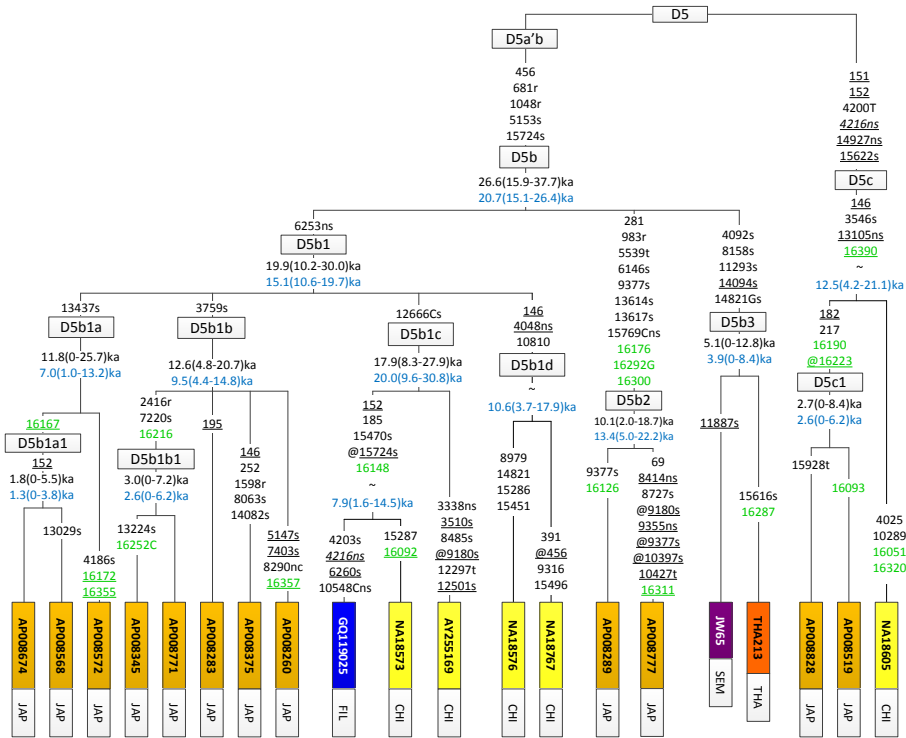


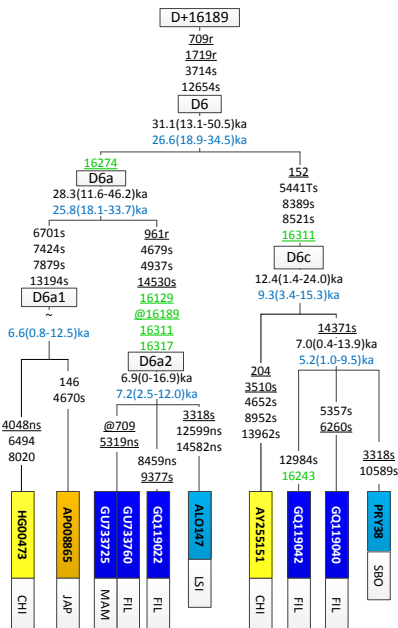
Figure 3.57 The tree of haplogroups D5b and D5c. Time estimates shown for clades are ML (in black) and averaged distance ( $p$ ; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan, SEM – Southeast Peninsular Malay, THA – Thailand)

### 3.26.3 Haplogroup D6

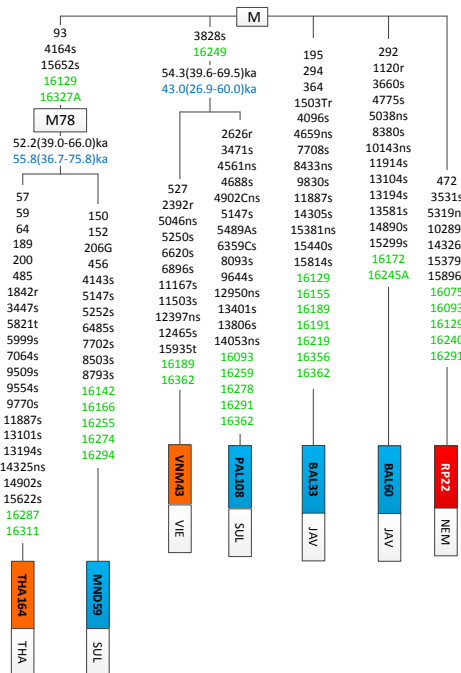
Haplogroup D6 dates to ~31 ka, and includes subclades D6a and D6c (Figure 3.58). D6 is more commonly found in SEA (more specifically in ISEA) compared to D5. The phylogeny does not suggest an MSEA origin, nor one in China, but includes Philippines haplotypes (including Mamanwa aboriginals) in both primary branches. This finding is not inconsistent with the M80 Philippines link above. **D6a** dates to ~28 ka and it further splits into D6a1 and D6a2. **D6a1** is seen in South China (Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004), and dates to ~7 ka (but dated with  $p$  only, due to time constraints). **D6a2** also dates to ~7 ka, and is seen in a single individual from Alor Island (Archaeogenetics Research Group, Huddersfield) and three from the Philippines (including a Mamanwa sequence; Tabbada *et al.*, 2010; Gunnarsdóttir *et al.*, 2011a). The HVS-I database shows that it is also seen in several Taiwanese aboriginal groups; all share the same HVS-I haplotype (matching the one seen in the tree here).

**D6c** dates to ~12 ka, with a basal type seen in South China (Kong *et al.*, 2003a) and a subclade including two haplotypes from the Philippines (Tabbada *et al.*, 2010) and one from Southern Borneo (Leeds Archaeogenetics Research Lab) ~7 ka. Given the errors on the age

estimates, both D6a2 and the Southeast Asian subclade of D6c are candidates as markers for the Austronesian dispersal.



**Figure 3.58** The tree of haplogroup D6. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FIL – Philippines, LSI – Lesser Sunda Islands, MAM – Philippines Mamanwa, SBO – South Borneo)



**Figure 3.59** The tree of haplogroups novel M\* and M78 (Zhang *et al.*, 2013). Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (SUL – Sulawesi, THA – Thailand, VIE – Vietnam)

### 3.27 Novel M\* Haplogroups

Several basal M\* haplogroups/lineages (this study) are found in SEA and not elsewhere in Asia (Figure 3.59). However, the first haplogroup has been recently named M78 (Zhang *et al.*, 2013) which is found in Thailand and Sulawesi, Indonesia (Archaeogenetics Research Group, Huddersfield) and dates to ~52 ka (~54 ka in Zhang *et al.*, 2013). According to Zhang *et al.* (2013), M78 is subdivided into two subclades, the first includes four lineages from the Austro-Asiatic-speakers Stieng tribe of Cambodia, where the Sulawesi lineage (MND59) shares a transition at np 5147 out of the 9 variants. The second subclade includes three previously unclassified lineages: two from Myanmar (JX289097, JX289130; Summerer *et al.*, 2014) and one from Tibet (HM030537; Kong *et al.*, 2011), where the Thai lineage (THA164) would share all 8 defining variants and nest within the same subclade (see Figure 4 in Zhang *et al.*, 2013).

The second haplogroup, defined by variants at nps 3828 and 16249, is seen in Vietnam and Sulawesi, Indonesia, dating to ~54 ka. They both show deep common ancestry between far-flung parts of the Sunda shelf, or close to it.

## 4 Results and Discussion: Haplogroup N

Haplogroup N is defined by transitions at nps 8701, 9540, 10398 and 10873, and a reversion at np 15301. It includes the major haplogroup R, which I deal with separately below. As shown in Figure 4.1, the other main branches in East and Southeast Asia are N9 (Tanaka *et al.*, 2004; Kong *et al.*, 2006; Metspalu *et al.*, 2006; Derenko *et al.*, 2007), N21, N22 (Macaulay *et al.*, 2005; Pierson *et al.*, 2006; Hill *et al.*, 2007; Tabbada *et al.*, 2010) and A (Tanaka *et al.*, 2004; Kong *et al.*, 2006; Derenko *et al.*, 2007; Achilli *et al.*, 2008), as well as three other smaller subclades N8, N10 and N11 (Kong *et al.*, 2011).

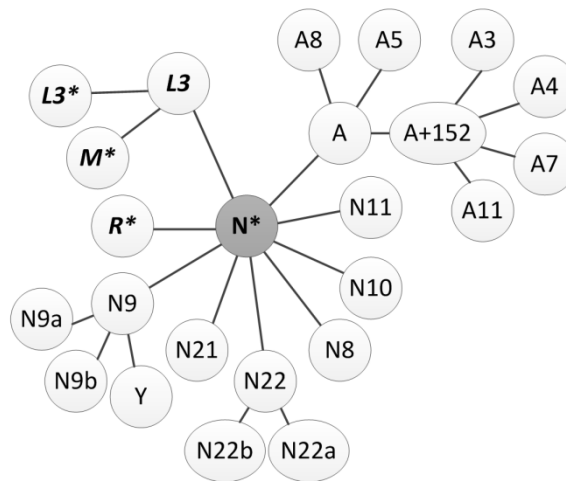


Figure 4.1 Schematic diagram of haplogroup N's major subclades present in Southeast Asia.

### 4.1 Haplogroup N9

Haplogroup N9, which is commonly distributed across Asia, is defined by a transition at np 5417 and dates to ~55 ka with three subclusters, N9a, N9b and Y. N9a has seven subclades, which are N9a1'3, N9a2'4'5, N9a6, N9a7, N9a8, N9a9 and N9a10. The N9 tree has includes a total of 125 complete sequences: 86 N9a, 19 N9b and 20 Y. Most of N9 subclades appear to be commonly found in China and Japan, with the deepest branches in China, which indicates it does not originate in Japan. The exception is N9a6, which is found mostly in ISEA. N9b has three subclades, N9b1, N9b2 and N9b3, which are reported in Japan only. Y can be divided into Y1 and Y2, where the basal lineages are found in China and Japan. Y2a is, however, found mostly in ISEA (Figure 4.2).

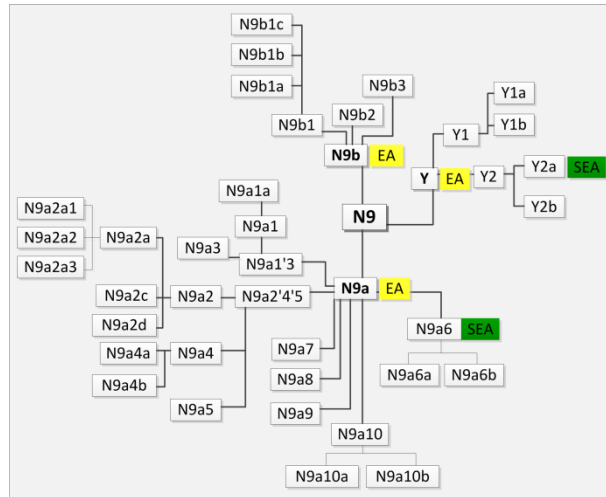


Figure 4.2 Schematic diagram of haplogroup N9 and its major subclades. (EA – East Asia, SEA – Southeast Asia)

#### 4.1.1 Haplogroup N9a

N9a has a divergence time of ~20 ka and it consists of seven subclades, N9a1'3, N9a2'4'5, N9a6, N9a7, N9a8, N9a9 and N9a10 (Figure 4.3).

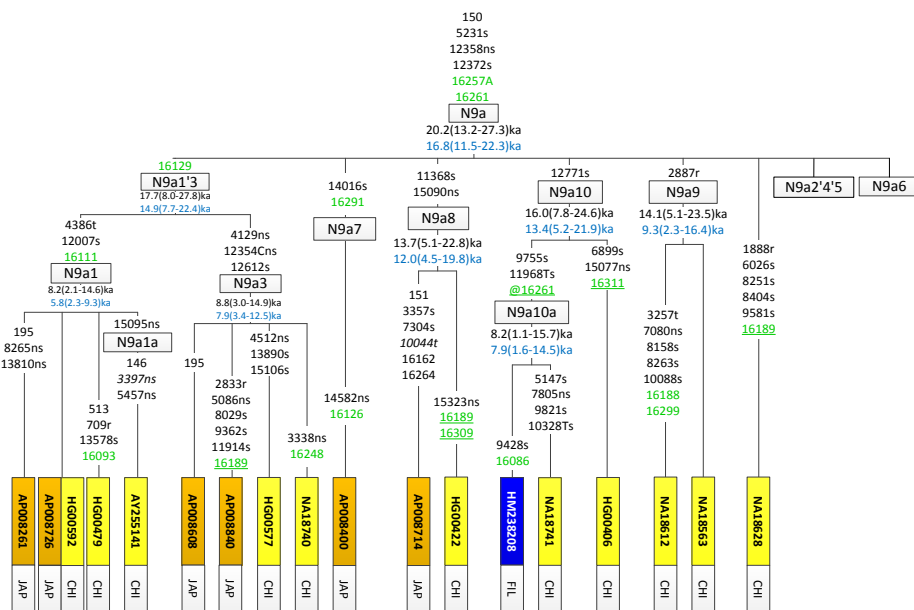
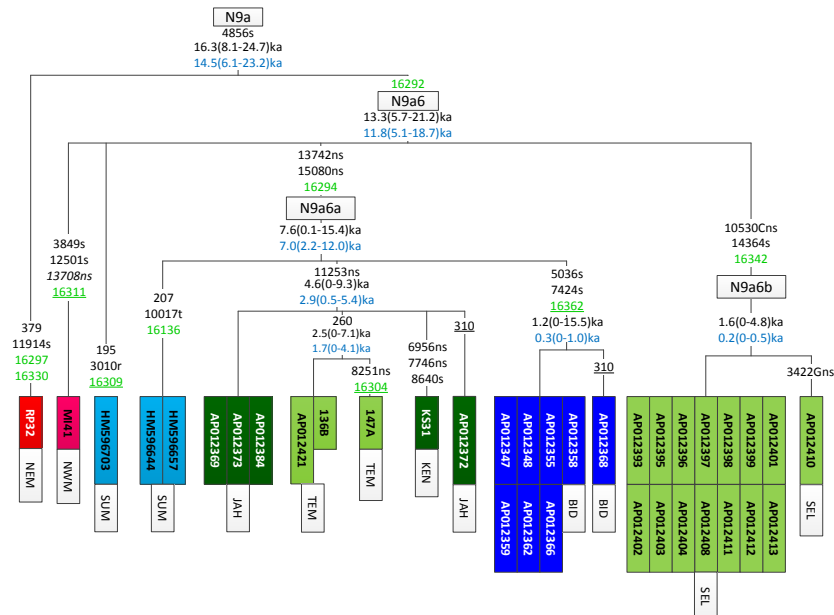


Figure 4.3 The tree of haplogroup N9a showing the subclades of N9a1'3, N9a2'4'5, N9a6, N9a7, N9a8, N9a9 and N9a10. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, FIL – Philippines, JAP – Japan)

N9a1'3 is weakly defined by a single, fast-evolving control region mutation at np 16129, dating to ~18 ka. Nested within this haplogroup are N9a1 and N9a3, both of which are found in East Asia. Subclade N9a1, dating to ~8 ka, and is found in south China (Kong *et*

*al.*, 2003b; Tanaka *et al.*, 2004; Zheng *et al.*, 2011). **N9a3** dates to ~9 ka, and has been reported in China and Japan (Tanaka *et al.*, 2004; Zheng *et al.*, 2011).

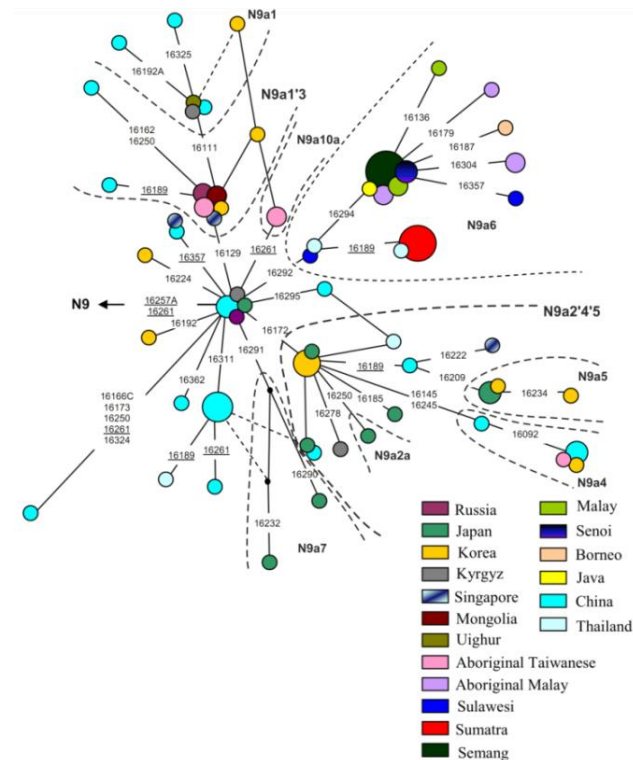
There is only a single, Japanese, individual belonging to **N9a7** represented in the tree (Tanaka *et al.*, 2004). **N9a8** is found in Japan (Tanaka *et al.*, 2004) and south China (Zheng *et al.*, 2011) dating to ~14 ka ( $\rho$  ~12 ka). **N9a9** is found in Beijing, China (Zheng *et al.*, 2011), and dates to ~14 ka ( $\rho$  ~9 ka). **N9a10** dates to ~16 ka with the deepest-branching lineage found in south China (Zheng *et al.*, 2011), and a subclade dating to ~8 ka seen in an Austronesian-speaking Filipino Ivatan (Loo *et al.*, 2011) and a northern Chinese individual (Zheng *et al.*, 2011). Ivatan Islanders are Austronesian speakers from Orchid Island and the Batanes archipelago located between Taiwan and the Philippines. They have been suggested to have a close cultural relationship with the Yami tribe in Taiwan, which is the only non-Formosan Austronesian speakers among Taiwan Aborigines (Blust, 1999), although Loo *et al.* (2011) found very little gene flow between Yami and Ivatan (i.e. as indicated by the limited sharing of mtDNA haplogroup B4a1a4 and O1a1\* MSY lineage). (N9a2'4'5 see Appendix E).



**Figure 4.4** The tree of haplogroup N9a6. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (JAH – Semang Jahai, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEL – Aboriginal Malay Seletar, TEM – Aboriginal Malay Temuan, SUM – Sumatra, BID – Sarawak Bidayuh)



**N9a6** is defined by transitions at nps 4856 and 16292, whereas a Kelantanese from northeast Peninsular Malaysia is found to possess only the transition at np 4856; the ‘pre-N9a6’ node (assuming it does not result from a reversion at np 16292) dates to ~16 ka (Figure 4.4). N9a6 dates to ~13 ka with derivatives found in all three main *Orang Asli* groups, Malay and across Indonesia. Subclade **N9a6a** dates to ~7.5 ka. This subclade is seen in Sumatra, Indonesia (Gunnarsdóttir *et al.*, 2011b), the *Orang Asli* Semang, and the Bidayuh of Sarawak (Jinam *et al.*, 2012). The Jahai (Jinam *et al.*, 2012) and Kensiuh (Semang, *Orang Asli*) shared the branch defined by a transition at np 11253 and its age has been estimated at ~5 ka. The Aboriginal Malay Temuans nested within this subclade, defined by a transition at np 260, dating to ~2.5 ka ( $p \sim 2$  ka); the Temuans here come from this study as well as one in Jinam *et al.* (2012). **N9a6b** is dated with a recent age of ~1.6 ka and it is found only in the Aboriginal Malay Seletar (Jinam *et al.*, 2012).



**Figure 4.5 Network of N9a\* and N9a6 from HVS-I data. Figure adapted from Hill (2005).**

N9a6 seems to have dispersed widely southwards from Mainland Southeast Asia into the Sunda during the Late Glacial period ~16 ka, and ultimately fissioned between all three *Orang Asli* groups and the Peninsular Malay, with some gene flow into Sumatra and ISEA. We can explore the distribution more comprehensively by turning to control-region data. In the HVS-I network in Figure 4.5, the root type of haplogroup N9a is found in China, Japan,

Russia and Kyrgyz of Central Asia, and derived types are commonly found in China, Japan and Korea (Hill, 2005). The subclade of N9a mtDNA types defined by a transition at np 16292, now labelled N9a6 in the complete-mtDNA tree, is found amongst the *Orang Asli* in the Senoi as well as Semang and Aboriginal Malays, Sumatran, Sulawesi, Borneo and Thailand. The network shows some overlap between Malay and all *Orang Asli* groups. The distribution of N9a6 suggests an origin in mainland Southeast Asia with spread into island Southeast Asia.

#### 4.1.2 Haplogroup N9b

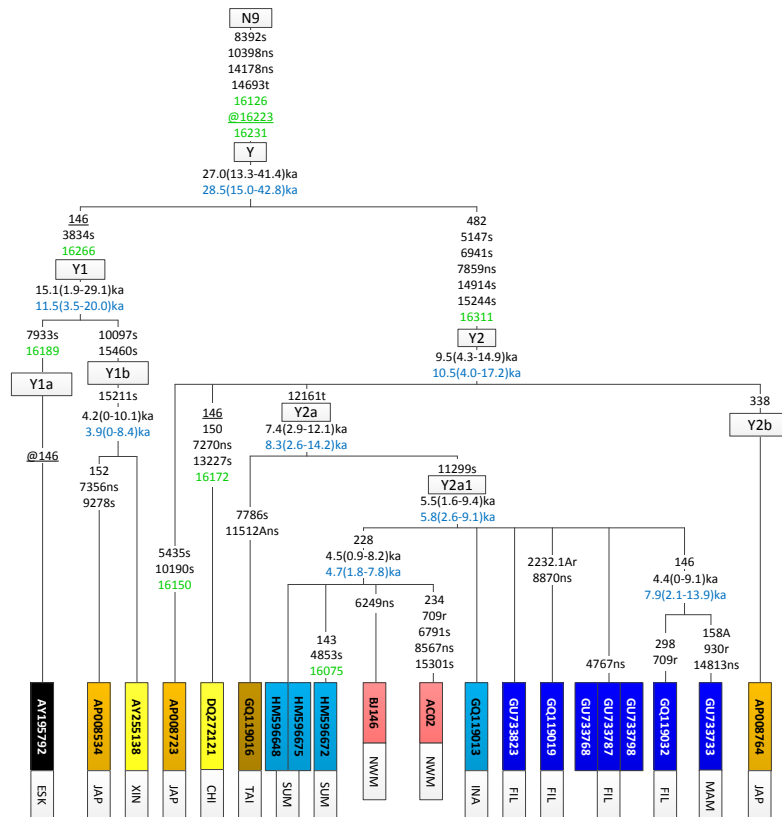
N9b dates to ~25 ka and the entire clade is found only in Japan (Tanaka *et al.*, 2004). Detailed description is available in Appendix E.

#### 4.1.3 Haplogroup Y

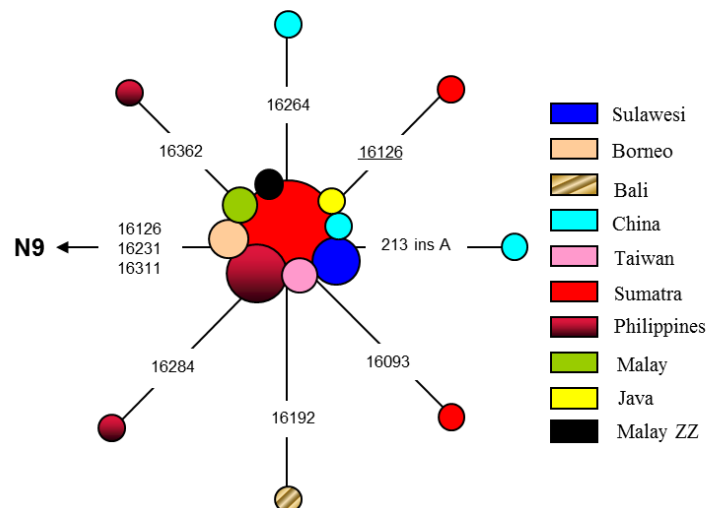
Haplogroup Y is further sub-divided into Y1 and Y2 (Figure 4.6). Y1 is seen in Japan and China, whereas Y2, especially Y2a, is commonly seen in mainland and island SEA. Haplogroup **Y1** dates to ~15 ka and divides into Y1a and Y1b. **Y1a** is defined by transitions at nps 7933 and 16189, which has been seen in an Eskimo reported by Mishmar *et al.* (2003). **Y1b** dates to ~4 ka and is seen in Xinjiang China (Kong *et al.*, 2003b) and Aichi Japan (Tanaka *et al.*, 2004). The HVS-I database confirms that it is largely restricted to East Asia, extending only as far south as South China. It is therefore likely to have an East Asian origin.

**Y2** dates to ~9.5 ka. Several basal lineages, including Y2b, are seen in Chongqing China (Kong *et al.*, 2006) and Aichi Japan (Tanaka *et al.*, 2004), again suggesting a likely origin in East Asia. It can be divided into Y2a and Y2b. **Y2a** dates to ~7 ka. There is a basal lineage in the Taiwanese Saisiat ethnic group (Tabbada *et al.*, 2010), possibly suggesting an origin amongst aboriginal Taiwanese. **Y2a1** dates to ~5.5 ka, and has spread widely in Island Southeast Asia, including both the Philippines (Gunnarsdóttir *et al.*, 2011a), and Indonesia (Tabbada *et al.*, 2010), but also to the Malay Peninsula. There are two subclades nested within Y2a1. The first subclade gained a transition at np 228, dated to ~4.5 ka, and is found in two Malay from Kedah and Perak, and Sumatra, Indonesia (Gunnarsdóttir *et al.*, 2011a). The second subclade has a transition at np 146, dated to ~4 ka, and is seen in the Philippines (Tabbada *et al.*, 2010) and the negrito Mamanwa (Gunnarsdóttir *et al.*, 2011a). **Y2b** is defined by a transition at np 338, which is only represented here by a sample from Aichi

Japan reported by Tanaka *et al.* (2004), again indicating an East Asian origin for Y2 and recent dispersal into the Sunda region, possibly via the Philippines.



**Figure 4.6** The tree of haplogroup Y. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, FIL – Philippines, INA – Indonesia, JAP – Japan, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, TAI – Taiwan, XIN – Xinjiang)



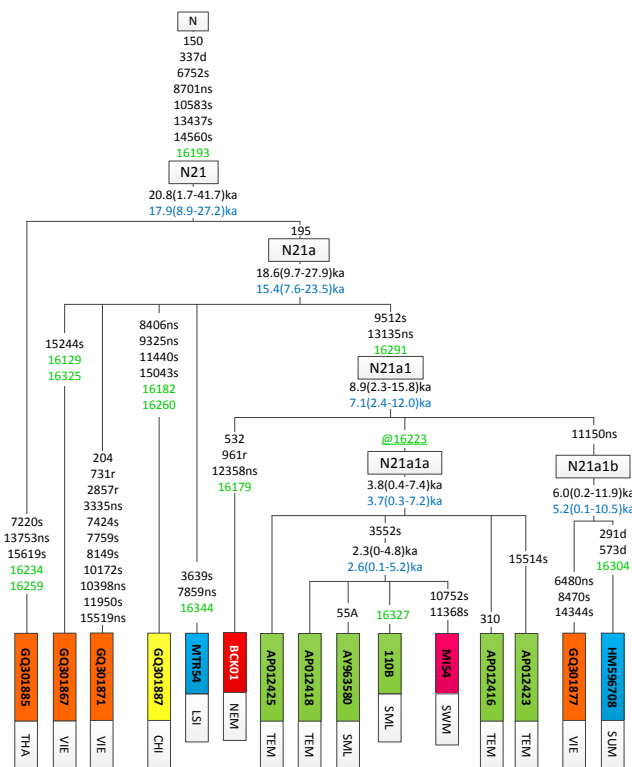
**Figure 4.7** HVS-I network of Y2. Figure adapted from Hill (2005).

Based on the published HVS-I data, we know that Y1 is commonly found in Japan, Korea, and the Kamchatka peninsula of North-East Russia with low frequency in China (Oota

*et al.*, 1995; Horai *et al.*, 1996; Lee *et al.*, 1997; Pfeiffer *et al.*, 1998; Schurr *et al.*, 1999; Yao *et al.*, 2002a, 2002b). Figure 4.7 shows the HVS-I network of Y2. Y2 HVS-I root types are found in Sumatra, Java, Sulawesi, Borneo, the Philippines, Taiwan, the Malay of Peninsular Malaysia, and one individual from Shanghai, and the derived ones are found in China, Sumatra, Bali and the Philippines. The Malay individuals overlap with those from ISEA at this level of resolution.

## 4.2 Haplogroup N21

N21 is basal within N (Macaulay *et al.*, 2005; Soares *et al.*, 2009) and the phylogeny is reconstructed by 15 complete mtDNA sequences. This haplogroup dates ~21 ka. The deeper lineages appear to be restricted to MSEA, detected in Thailand, Vietnam (Peng *et al.*, 2010) and Yunnan China (Kong *et al.*, 2011). The recent expansion events brought the lineages into Peninsular Malaysia and the Aboriginal Malays, and also Indonesia (Figure 4.8).

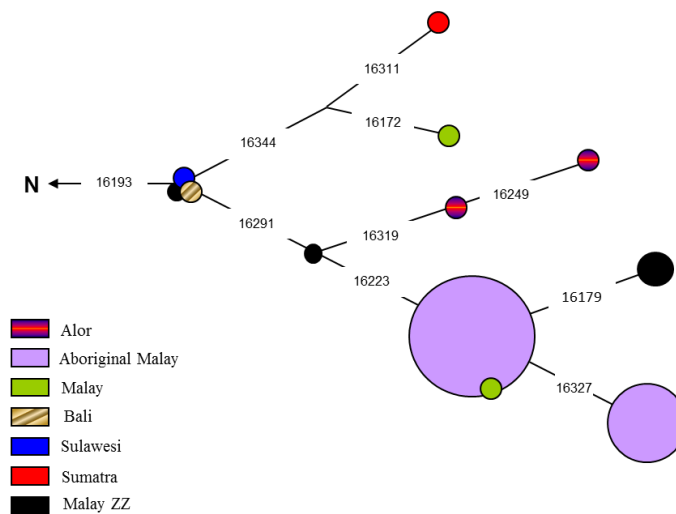


**Figure 4.8** The tree of haplogroup N21. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, LSI – Lesser Sunda Islands, NEM – Northeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUM – Sumatra, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan, THA – Thailand, VIE – Vietnam)

Subclade **N21a** dates to ~ 19 ka, where it is seen in the Cham individuals of Bin Thuan Vietnam (Peng *et al.*, 2010), Chinese in Lijiang Yunnan (Kong *et al.*, 2011), and Mataram in

Lombok Indonesia (Archaeogenetics Research Group, Huddersfield). **N21a1** dates to ~9 ka and is seen in Aboriginal Malays, a Malay from Kelantan (this study), as well as singletons in Vietnam and Sumatra. N21a1 is divided into N21a1a and N21a1b. The Aboriginal Malays and a southwest Malay are nested within clade **N21a1a**, which dates to ~4 ka, including six Aboriginal Malays: Semelai (Macaulay *et al.*, 2005), Temuan (Jinam *et al.*, 2012) and a Malay Minangkabau from Southwest Peninsular Malaysia (this study). The Minangkabau have an ancestry that can be traced very recently to West Sumatra in Indonesia. **N21a1b** dates to ~6 ka and is found in two individuals from Bin Thuan of Vietnam (Peng *et al.*, 2010) and Sumatra, Indonesia (Gunnarsdóttir *et al.*, 2011b).

In Hill *et al.* (2006), the root type of N21 (when using HVS-I data) is found in Bali, Sulawesi (Palu), and Malaysia. One of the branches nested within this root type is characterised by a transition at np 16344 and is found in Sumatra (Palembang) and Peninsular Malaysia (Zainuddin and Goodwin, 2004). The second branch is defined by a transition at np 16291 is now identified as N21a1 and is found in Alor, the Semelai, the Temuan and Peninsular Malaysia (Figure 4.9).



**Figure 4.9** HVS-I network of N21. Denotation “Malay” is the Malay data used in Hill *et al.* (2006), “Malay ZZ” is the new data from Zafarina Zainuddin (personal communication). Figure adapted from Hill (2005).

### 4.3 Haplogroup N22

In Figure 4.10, N22 is basal to N and the tree is reconstructed using 10 complete mtDNA sequences. N22 is defined by transitions at nps 150, 942, 7158, 9254, 11365, 16168 and 16249 (Macaulay *et al.*, 2005). It dates to ~29 ka and the tree indicates it as a deep Sunda haplogroup, since it is only seen in SEA, in particular the Aboriginal Malays, Malaysia,

Indonesia although also the Philippines. N22 diverged into subclades N22a and N22b – the latter being a newly defined haplogroup.

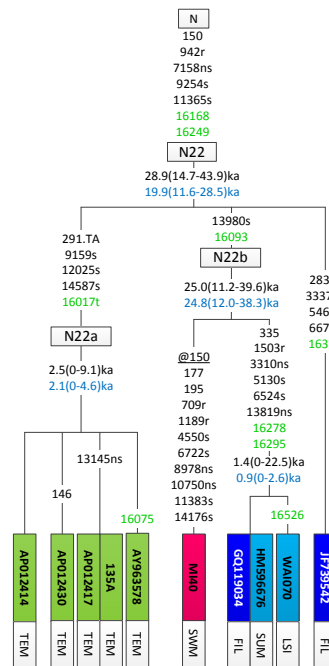


Figure 4.10 The tree of haplogroup N22. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (FIL – Philippines, LSI – Lesser Sunda Islands, SUM – Sumatra, SWM – Southwest Peninsular Malay, TEM – Aboriginal Malay Temuan)

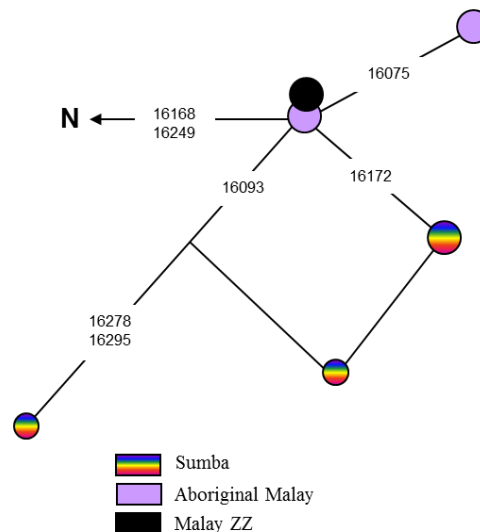


Figure 4.11 HVS-I network of N22. Figure adapted from Hill (2005).

One non-negrito Philippine Cuyonin tribe (Scholes *et al.*, 2011) was found to be basal within N22. The older subclade, **N22b**, dating to ~25 ka and is detected in one Minangkabau sequence from Negeri Sembilan which has accumulated a high number of private mutations on its tip. N22b is also seen in Indonesia (one Sumba from this study and one in Sumatra

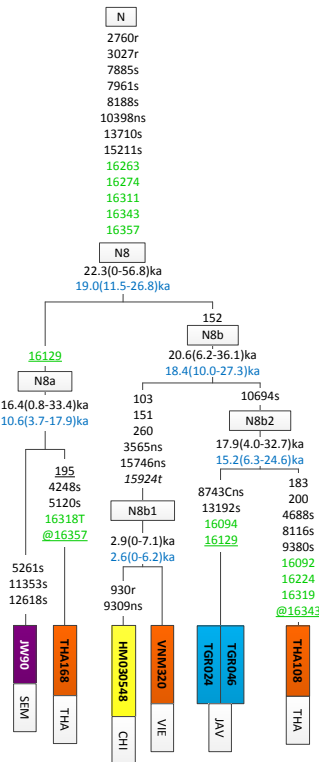
(Gunnarsdóttir *et al.*, 2011b) and the Philippines (Tabbada *et al.*, 2010), which formed a subclade dating to ~1.4 ka. **N22a** is much more recent, with a ~2.5 ka divergence time, and highly localised in the Temuan population, which has been subjected to founder effect and genetic drift (data from this study, Macaulay *et al.*, 2005 and Jinam *et al.*, 2012).

In Figure 4.11, the network for HVS-I *Orang Asli* data shows haplogroup N22 is seen only in the Aboriginal Malay Temuan characterised by nps 16168 and 16249. The root type of N22 is found in the Temuan and Peninsular Malaysia, with its derived types found in Temuan and Sumba. The network shows a sharing of lineage type between Malay and Aboriginal Malays.

#### 4.4 Haplogroup N8

N8 is a basal haplogroup that has undergone high drift as seen with the long internal branch resulting in a divergence time of ~22 ka (Figure 4.12). The N8 tree is reconstructed by 7 complete mtDNA sequences. It is characterised by nps 2760, 3027, 7885, 7961, 8188, 10398, 13710, 15211, 16263, 16274, 16311, 16343 and 16357 (Kong *et al.*, 2011). N8 is mostly seen in Mainland and Island SEA, like N22, is largely a Sunda lineage. The deep lineages are found in northern Mainland SEA especially in North Thailand and North Vietnam, which appears to have originated in northern Mainland SEA and then expanded into Island SEA. The HVS-I data (it was called haplogroup N6 in Mormina, 2007) showed that N8 is found in West Indonesia, Peninsular Malaysia and MSEA.

N8 is divided into N8a and N8b; both are annotated here for the first time. **N8a** is defined by a transition at np 16129, with an estimated date of ~16 ka. N8a is seen in Jawa Malay of Southeast Peninsular Malaysia (this study) and one Thai individual (Archaeogenetics Research Group, Huddersfield). **N8b** is characterised by np 152, dating to ~21 ka, which can be further divided into two subclades, N8b1 and N8b2. **N8b1** is defined by transitions at nps 103, 151, 260, 3565, 15746 and 15924, and dated ~3 ka. N8b1 is detected in Guizhou China (Kong *et al.*, 2011) and Vietnam (Archaeogenetics Research Group, Huddersfield). **N8b2** is defined by a transition at np 10694 and estimated at ~18 ka. It is found on the Tengger Island in Java Timur Indonesia as well as Thailand (Archaeogenetics Research Group, Huddersfield; Hill *et al.*, 2006; Mormina, 2007).



**Figure 4.12** The tree of haplogroup N8. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAV – Java, SEM – Southeast Peninsular Malay, THA – Thailand, VIE – Vietnam)

## 4.5 Haplogroup N10

N10 (Kong *et al.*, 2011) dates to ~63ka – approximately the age of haplogroup N itself (Figure 4.13). The N10 tree includes six complete mtDNA sequences from South China, and one individual each from Northeast Peninsular Malaysia and South Borneo, which may suggest an ancient origin in southern China/Mainland SEA.

In Figure 4.13, N10 diverges into two subclades, N10a and N10b. **N10a** dates to ~55 ka. N10a is further divided into N10a1 and N10a2. **N10a1** dates to ~2.5 ka. N10a1 is seen, so far, only in SEA (Northeast Peninsular Malaysia: this study) and South Borneo, Indonesia (Archaeogenetics Research Group, Huddersfield). **N10a2** has an estimated age of ~29 ka. It is seen in Yunnan China (Kong *et al.*, 2011) before spreading into the north in Beijing (Zheng *et al.*, 2011) and Xinjiang China (Kong *et al.*, 2011) ~7 ka. Lastly, **N10b** is seen in a Guangdong individual, where the same HVS-I haplotypes are present in southern Chinese: Shanghai and Jiangsu (Kong *et al.*, 2011). The long internal branches in the phylogeny suggest founder effect and genetic drift, again indicating N10 have an origin in South China/Mainland SEA.



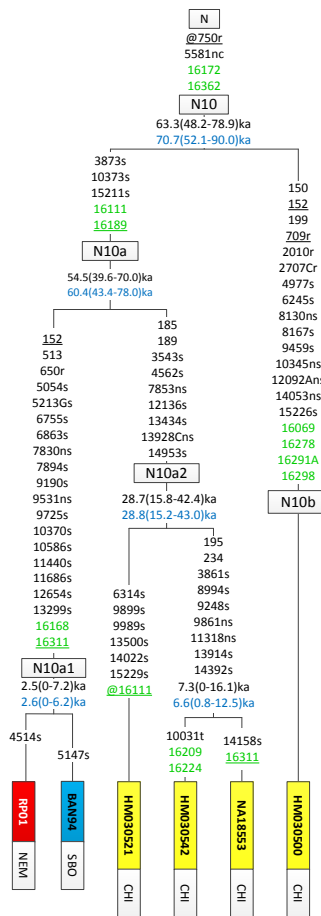


Figure 4.13 The tree of haplogroup N10. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NEM – Northeast Peninsular Malay, SBO – South Borneo)

## 4.6 Haplogroup N11

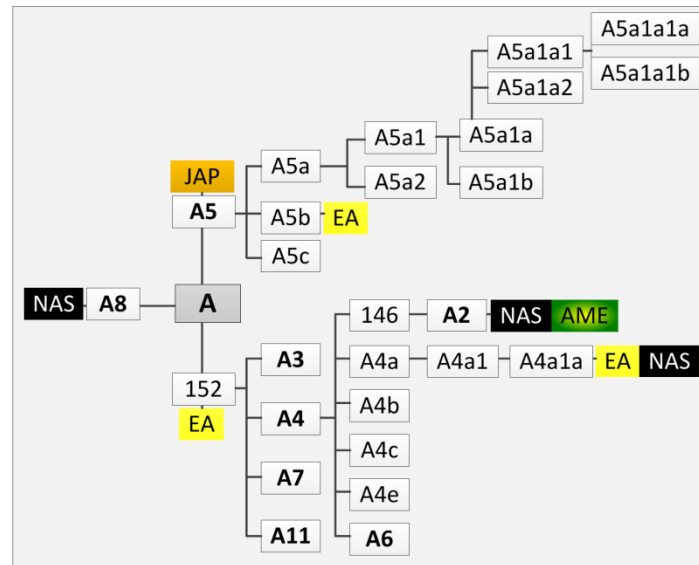
N11 is an ancient basal haplogroup defined by a transversion at np 11581A that dates to ~68 ka. The tree of N11 is built from 18 complete mtDNA sequences. One subclade, N11a, is found in Inner Mongolia, China and Tibet, and the other, N11b, in the negrito Mamanwa of the Philippines.

**N11a** dates to ~20 ka (Figure 4.14). It looks to have an origin in northeast Asia. It is detected in Inner Mongolia of China (Kong *et al.*, 2003), and further diverged into **N11a1**, which dates to ~11 ka. Nested within N11a1 are individuals from Sichuan, China and Naqu, Tibet (Kong *et al.*, 2011).

**N11b** dates to ~7 ka ( $\rho$  ~10 ka). This deeply-diverged subclade appears to belong exclusively to the negrito Mamanwa of the Philippines (Gunnarsdóttir *et al.*, 2011a). N11b



## 4.7 Haplogroup A



**Figure 4.15** Schematic diagram of haplogroup A and its major subclades. (AME – America, EA – East Asia, NAS – North Asia)

Haplogroup A is one of the more frequent East Asian haplogroups, reaching very high frequencies in northeast Asia (Kivisild *et al.*, 2002) and dates to ~27 ka. The phylogeny includes 78 complete mtDNA sequences: 49 A5, 21 A4, and 8 from minor subclades. Haplogroup A has three subclades, A5, A8, and A+152, where the latter includes A2, A3, A4, A7 and A11.

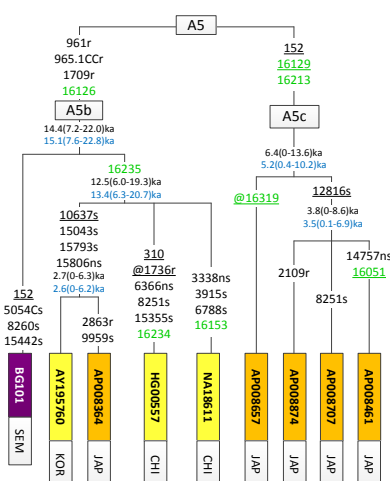
The use of np 152 and np 146 in Phylotree (van Oven and Kayser, 2009) to define subclades A+152 and A+146 are very problematic. These nucleotide positions are fast-evolving sites identified by Soares *et al.* (2009) and their evolutionary histories are sometimes difficult to track. For example, nesting in A4 of Native Americans within Japanese lineage is probably an artefact. They are unlikely to be true clades but retain here due to its presence in Phylotree.

Figure 4.15 shows the schematic diagram of phylogeny A with its distribution. The whole haplogroup is commonly found across China and Japan. Analyses carried out with HVS-I data have shown that haplogroup A is commonly reported across China, Japan and Korea (Lee *et al.*, 1997; Nishimaki *et al.*, 1999; Kivisild *et al.*, 2002; Yao *et al.*, 2002a; Yao *et al.*, 2002b). A4 and A5 are the most diversified subclades, while A2, a subclade of A4, is found in northeast Asia and America. In the whole-mtDNA trees, A5 has deepest roots in China, while the roots for A4 lie in Japan. So far the reported haplogroup A lineage is limited

from SEA except for a Southeast Peninsular Malay from Johor Malaysia, belonging to A5b (this study).

**A5** dates to ~21 ka, and consists of A5a, A5b and A5c. **A5a** is mainly a Japanese haplogroup (Tanaka *et al.*, 2004), dates ~9 ka and has diverged into A5a1 (~8 ka), A5a2 (dates to ~2 ka, restricted to Tokyo and Aichi) and A5a3 (6 ka, restricted to Tokyo and Aichi). Detailed description is available in Appendix E.

In Figure 4.16, **A5b** dates to ~14 ka. A single example is found in a Bugis Malay in Southeast Peninsular Malaysia (this study). The main subclade nested within A5b is defined by a transition at np 16235, which is found in China around ~13 ka. A further nested subclade dates to ~3 ka and is found in Japan (Tanaka *et al.*, 2004) and Korea (Mishmar *et al.*, 2003). **A5c**, like A5a, is an entirely Japanese clade reported by Tanaka *et al.* (2004), dating to ~6 ka. The main nested subclade dates to ~4 ka and is found in Aichi, Chiba and Tokyo.



**Figure 4.16** The tree of haplogroup A5b and A5c. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, KOR – Korea, SEM – Southeast Peninsular Malay)

One **A8** sample is represented in the tree and found in Siberian Russia by Starikovskaya *et al.* (2005) (Figure 4.17). In control-region data, it is indeed most frequently seen in the Kamchatka Peninsula of northeast Siberia, but has also been seen, with low variation, in Mongolia, Japan and Kazakhstan (Kolman *et al.*, 1996; Horai *et al.*, 1996; Comas *et al.*, 1998; Schurr *et al.*, 1999).

Haplogroup **A+152** (retained here due to its presence in PhyloTree, but this is extremely unlikely to represent a true clade since np 152 is an extremely fast-evolving site: Soares *et al.*, 2009) dates to ~21 ka, and includes A3, A4, A7 and A11, of which **A4** is by far the most frequent and widespread. This notwithstanding, the tree suggests an origin in China

alongside the other haplogroup A subclades, with some lineages spreading into Japan and northeast Asia, and ultimately (as haplogroup A2) into the Americas. **A4a1** dates to ~11 ka and is seen in Mongolia (Hartmann *et al.*, 2009) and Japan (Tanaka *et al.*, 2004). **A4b** and **A4c1** are represented by an individual from Siberia, Russia (Starikovskaya *et al.*, 2005) and one Chinese from Naxi China (Hartmann *et al.*, 2009) respectively.

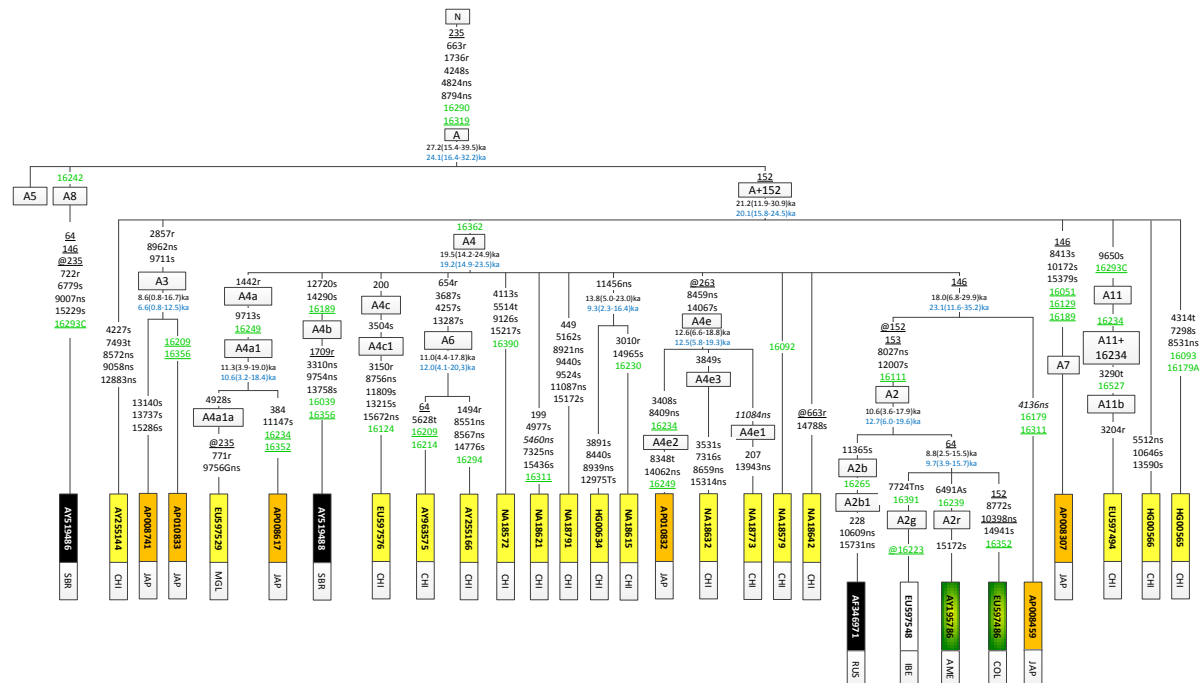
**A4+146** is again present in PhyloTree but extremely implausible as a true clade since np 146 is another extremely fast-evolving site, is the putative pre-A2 node nested within A4 and hence, the basal lineage possibly present in Japan (Tanaka *et al.*, 2004). **A2** dates to ~11 ka, where it spreads north into Chukchi Russia (Ingman *et al.*, 2000). This Siberian individual is the only A2b1 represented in this tree. Lastly, a subclade nested within A2 is divided into A2g and A2r, and dating to ~9 ka. This subclade is largely a Native American lineage (Hartmann *et al.*, 2009). One sample represents each haplogroup in this subclade (Figure 4.17), **A2r** is seen in a Native American (Mishmar *et al.*, 2003), **A2g** surprisingly by a sample of Iberia ancestry as noted by Hartmann *et al.*, (2009), and one undefined lineage from Columbia (Hartmann *et al.*, 2009).

**A3** dates to ~9 ka; it is seen only in Japanese from Tokyo (Nohira *et al.*, 2010) and Aichi (Tanaka *et al.*, 2004). **A6** dates to ~11 ka and seen in individuals reported by Kong *et al.* (2003b) from Hubei China and Macaulay *et al.* (2005) in Tujia China. **A4+11456** is a subclade dated to ~14 ka and seen in south China (Zheng *et al.*, 2011). **A4e** dates to ~13 ka, and is observed in Japan and China. Although each subclade is only represented by one sample, they are from Beijing China for A4e1 and A4e3 (Zheng *et al.*, 2011), and A4e2 in an Ehime Japanese (Nohira *et al.*, 2010). **A7** is represented here by a sample from Japan (Tanaka *et al.*, 2004). Similarly, **A11b** is seen here from a Naxi China sample (Hartmann *et al.*, 2009).

Haplogroup A dates to pre-LGM ~27 ka. Haplogroup A5 and its subclades are highly diversified and commonly found in Japan. A5a started to spread during the early Holocene ~21 ka throughout Japan, which has only one lineage from Inner Mongolia found nested within A5a1a1. A5b dates to the late Pleistocene ~14 ka. The lineage found in Southeast Peninsular Malaysia may suggest a relict that survived since the beginning of sea-level rise ~15 ka in the Sunda region. The bigger subclade of A5b has basal lineages found in South and North China, which appears to spread ~13 ka from China into Japan and Korea recently ~3 ka. A5c as mentioned before has arrived in Japan during the early Neolithic ~6 ka to ~4 ka. Similar to A5, A4 dates to the end of LGM ~19 ka, but A4 and its subclades are

commonly found in China, largely arrived by the early Holocene. Considering the fact that A5b has a date older than A5a and A5c, and that A5b is present in Peninsular Malaysia and South China, A5 in overall might have a pre-LGM southern origin before expanded into North Asia after early Holocene.

A2, together with B2, C1, and D1, are the four main “pan-American” haplogroups that dispersed into the New World (Achilli *et al.*, 2008; Bandelt *et al.*, 2008). A2 dates to the early Holocene, ~11 ka in Russia, and a subclade nested within, dating to ~9 ka, is seen in the America continent in the Native American, Colombian, and surprisingly, a single instance of Iberian ancestry (Hartmann *et al.*, 2009).



**Figure 4.17** The tree of haplogroup A8 and A+152. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (SBR – Siberian Russia CHI – China, JAP – Japan, MGL – Inner Mongolia, China, RUS – Russia, IBE – Iberian Peninsula, AME – America, COL – Columbia)

## 5 Results and Discussion: Haplogroup R

Haplogroup R branches from the root of N with transitions at nps 12705 and 16223, dated by ML to ~68 ka. Several major Asian R haplogroups found in this study are shown in Figure 5.1, they include: B4, B5, R11'B6, R12'21, R22, R9, and P, and other rarely seen haplogroups in Asia including R6, R7, R23, R30, and U.

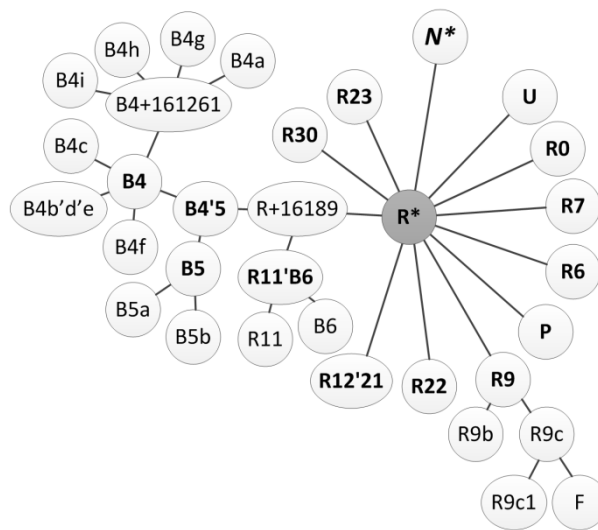


Figure 5.1 Schematic diagram of haplogroup R's major subclades present in Southeast Asia.

### 5.1 Haplogroup B

Haplogroup B nested within R+T16189C, which can be divided into B4'5 and R11'B6. B4'5 is one of the most common haplogroups in ISEA, consisting of haplogroups B4 and B5. Haplogroup B4'5 is defined by a 9 base-pair deletion at nps 8281-8289 in the COII/tRNA<sup>Lys</sup> intergenic region and a transition at np 16189 (Soares *et al.*, 2007; Hartmann *et al.*, 2009; Kong *et al.*, 2011; Soares *et al.*, 2011), dates by ML to ~62 ka, and B4 ~51 ka (estimated at ~44 ka in Soares *et al.*, 2011) (Figure 5.3). Haplogroup R11'B6 is divided into R11 and B6, where R11 is mainly restricted to China, and B6 is widely distributed in SEA. B6 shares the defining node of B (R+T16189C) with B4'5 that is hypervariable, they are most likely phylogenetically unrelated.

The phylogeny of B4 is reconstructed from 300 complete mtDNA sequences: 163 B4a, 44 B4b, 68 B4c, 9 B4d and 16 in minor subclades. The schematic diagram below (Figure 5.2)

shows B4 and its major subclades have deep ancestral roots and are widespread across East Asia and SEA, with more recent dispersals of some subclades into the Pacific (Lum *et al.*, 1998; Pfeiffer *et al.*, 1998; Kivisild *et al.*, 2002; Yao *et al.*, 2002b). Some of the subclades have been subjected to founder effect during the secondary expansions. The subclades more commonly seen in SEA are B4a1a, B4a1c4, B4a2a, B4g1a, B4b1a2, B4c1b2a2 and B4c2. B4a1c2 is seen in north Eurasia (Russia and Eskimo), and B4a1a1a and B4a1a1b have risen to high frequency and are almost fixed in Near Oceania (Soares *et al.*, 2011). Overall, the pattern suggests an ancient ancestry in East Asia and dispersal into SEA after the LGM (Soares *et al.*, 2011).

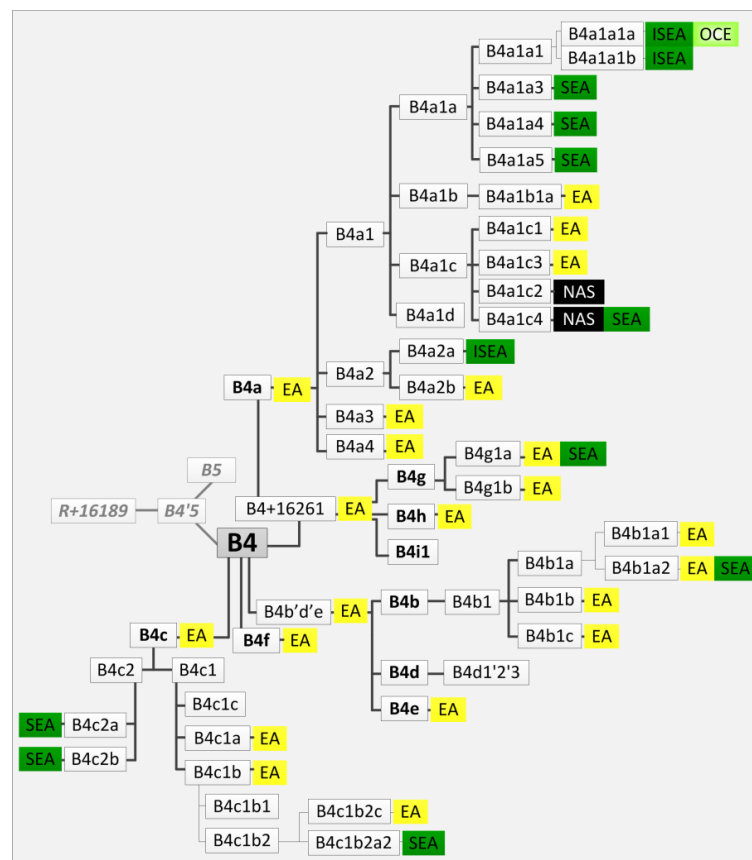


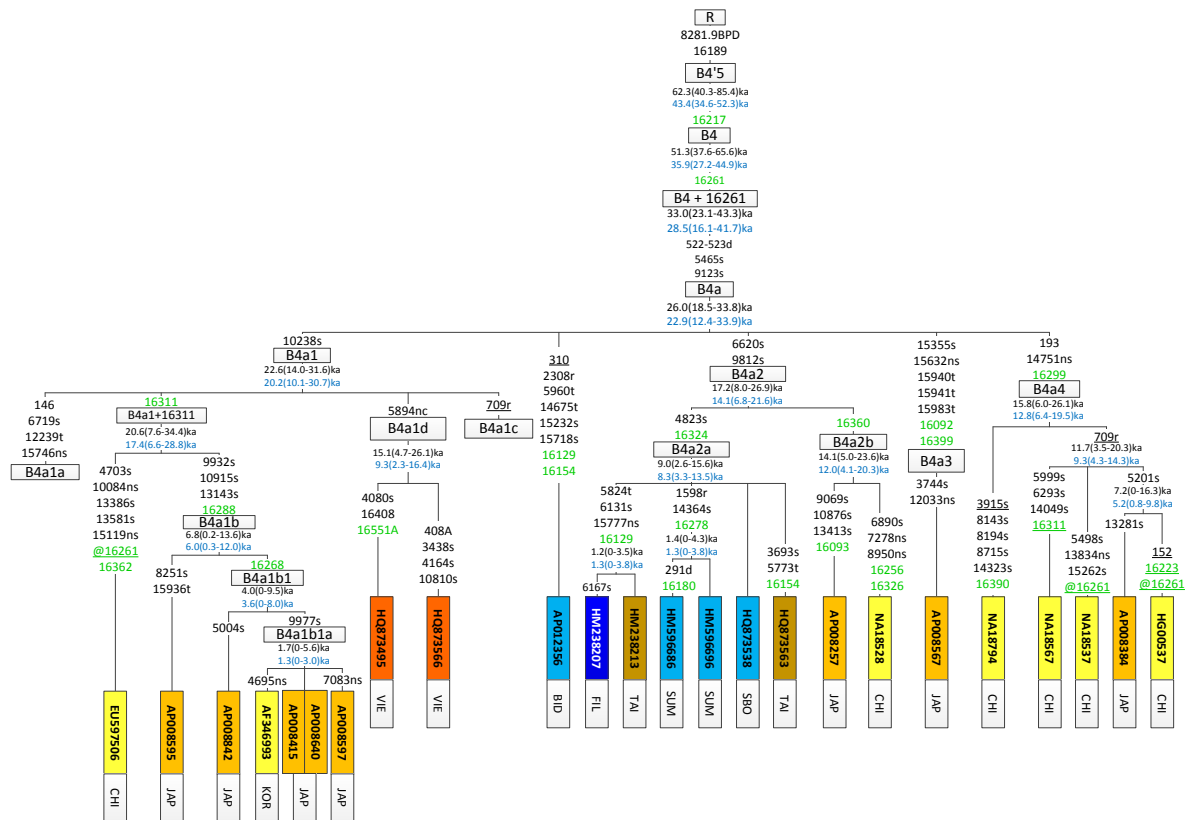
Figure 5.2 Schematic diagram of haplogroup B4 and its major subclades distribution. (EA – East Asia, NA – North Asia, NEU – North Eurasia, SEA – Southeast Asia and NO – Near Oceania)

### 5.1.1 Haplogroup B4+C16261T

**B4+C16261T** dates to ~33 ka, and is divided into B4a, B4g, B4h and B4i (Figure 5.3). B4a dates to ~26 ka and has four subclades. **B4a2** dates to ~17 ka. **B4a2a** dates to ~9 ka and is found mainly in Taiwan (Trejaut *et al.*, 2005), and much more rarely in the Philippines, Sumatra and Banjarmasin in southern Borneo, Indonesia. There are two subclades within



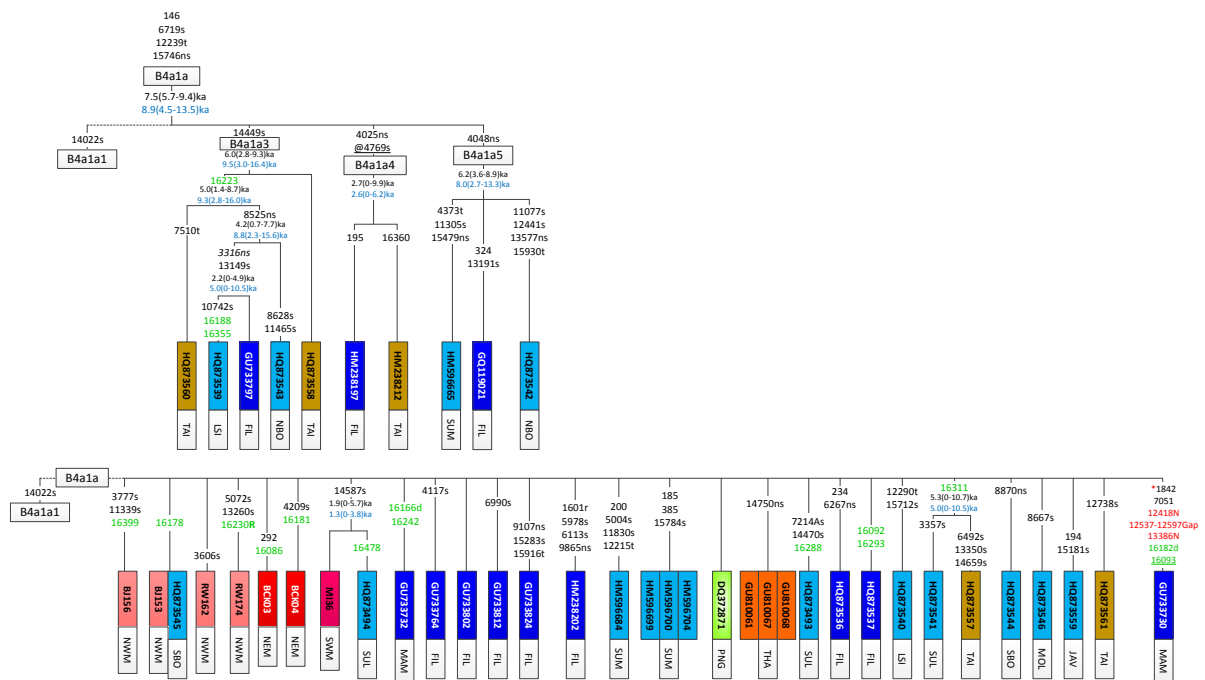
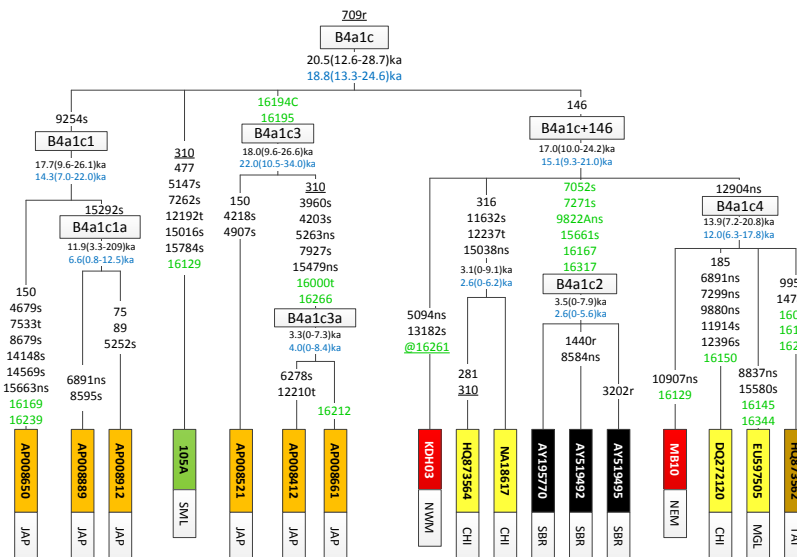
B4a2a; the first, dating to ~1.2 ka, is seen in the Ami people in Taiwan and the Philippines Batan archipelago (Trejaut *et al.*, 2005; Loo *et al.*, 2011) and the second in Sumatra, Indonesia, dating to ~1.4 ka (Gunnarsdóttir *et al.*, 2011b; Soares *et al.*, 2011). The HVS-I data shows B4a2 is found mainly among Taiwanese aboriginals, and it is not represented in MSEA (Mormina, 2007). **B4a2b** dates to ~14 ka, and is seen in Gifu, Japan (Tanaka *et al.*, 2004) and north China (Zheng *et al.*, 2011).



**Figure 5.3** The tree shows haplogroup B4a excluding B4a1a and B4a1c. Time estimates shown for the clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. (CHI – China, JAP – Japan, VIE – Vietnam, BID – Sarawak Bidayuh, FIL – Philippines, KOR – Korea, TAI – Taiwan, SUM – Sumatra, SBO – South Borneo)

**B4a3** is represented here by one Japanese Aichi individual reported by Tanaka *et al.* (2004). **B4a4** dates to ~16 and is seen in north China (Zheng *et al.*, 2011). A further mutation at np 709 defines a subclade dated to ~12 ka, it then obtained np 5201 which dates to ~7 ka, seen in south China (Zheng *et al.*, 2011) and Tokyo Japan (Tanaka *et al.*, 2004).

**B4a1** has four subclades, B4a1a, B4a1+16311 (which includes B4a1b), B4a1c and B4a1d. **B4a1b** dates to ~7 ka and is restricted to China and Japan. **B4a1b1** dates to ~4 ka, with a nested subclade **B4a1b1a**, dating to ~2 ka, and is seen in Japan (Tanaka *et al.*, 2004) and Korea (Ingman *et al.*, 2000). On the other hand, **B4a1d** is seen in two individuals from Vietnam (Soares *et al.*, 2011) and its age is estimated at ~15 ka.



**B4a1c** dates to ~20.5 ka and its three subclades, B4a1c1, B4a1c3 and B4a1c+146, are seen in East Asia, North Asia, Peninsular Malaysia with, interestingly, a fourth basal lineage seen in the Aboriginal Malay Semelai (Figure 5.4). **B4a1c1** and **B4a1c3** both date to ~18 ka

and are seen only in Japan (Tanaka *et al.*, 2004). **B4a1c with np T146C** dates to ~17 ka, and includes B4a1c2 and B4a1c4, and a subclade in China. B4a1c+146 is seen in China, Siberian Russia, Taiwan and Peninsular Malaysia. **B4a1c2** dates to ~3.5 ka, and is seen in two Russian Siberians (Starikovskaya *et al.*, 2005) and a Southeastern Siberia Eskimo (Mishmar *et al.*, 2003). **B4a1c4** dates to ~14 ka and is observed in southern China (Guizhou) (Kong *et al.*, 2006), Mongolia (Hartmann *et al.*, 2009), Taiwan (Soares *et al.*, 2011) and Peninsular Malaysia (this study).

**B4a1a** dates to ~8 ka (Figure 5.5). Basal lineages are commonly seen throughout SEA, including the Philippines, Indonesia (Gunnarsdóttir *et al.*, 2011b; Soares *et al.*, 2011), Peninsular Malaysia (this study), Thailand (Pradutkanchana, Ishida and Kimura, 2010), and Aboriginal Taiwan Ami and Tsou (Soares *et al.*, 2011), as well as Papua New Guinea (Pierson *et al.*, 2006). This starburst pattern points to a dramatic expansions across the region, centred on ISEA, in the early Holocene, similar to that seen in haplogroup E (Soares *et al.*, 2008). It has four major subclades: B4a1a1, B4a1a3, B4a1a4 and B4a1a5, with two other unnamed subclades.

**B4a1a3** dates to ~6 ka. A subclade with np 16223 dates to ~5 ka and a single basal lineage for each are seen in Taiwanese Ami and Siraya tribes, respectively (Soares *et al.*, 2011). A further subclade dates to ~4 ka, and is seen in Kota Kinabalu Malaysia (Soares *et al.*, 2011), with a further nested subclade nested within, dating to ~2 ka, seen in the Philippines Manabo (Gunnarsdóttir *et al.*, 2011a) and Sumba Indonesia (Soares *et al.*, 2011). This nesting relationship might imply (although with very few samples) dispersal from Taiwan into ISEA ~4 ka, although in the context of an earlier dispersal in the reverse direction, given the major earlier radiation in ISEA of basal B4a1a lineages.

**B4a1a4** dates to ~3 ka and seen in the Yami of Taiwan and the Philippine Ivatan (Loo *et al.*, 2011). Also see N9a10. **B4a1a5** dates to ~6 ka, with single instances seen in Sumatra, Indonesia (Gunnarsdóttir *et al.*, 2011b), the Philippines (Tabbada *et al.*, 2010) and Kota Kinabalu, Malaysia (Soares *et al.*, 2011).

**B4a1a1** (Figure 5.6) dates to ~7 ka and nested within are B4a1a1a, B4a1a1b, two small unnamed subclades and several paraphyletic lineages. **B4a1a1b** dates to ~4 ka and is seen in Kapingamarangi and Majuro Atolls, Micronesia only (Pierson *et al.*, 2006). The first unnamed subclade defined by a transition at np 6905 and dated to ~3 ka, while the second unnamed subclade is defined by a transition at np 16129 with a date of ~5 ka. The majority of

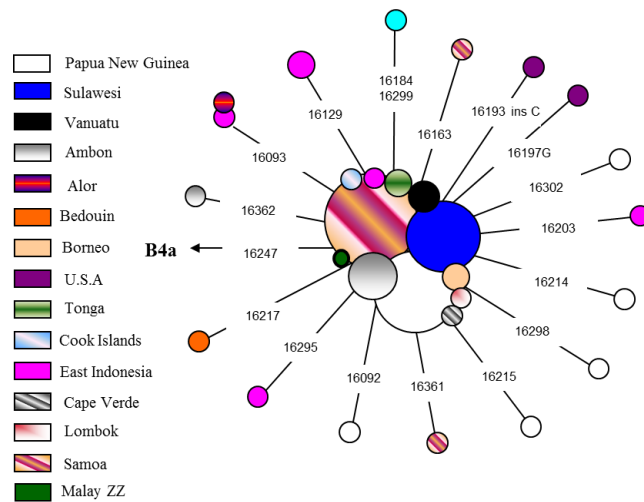
[illegible]

**B4a1a1a** dates to ~6 ka, and includes B4a1a1a1, B4a1a1a4, and other unnamed subclades (Figure 5.7). **B4a1a1a1** dates to ~5.5 ka and is seen only in Near Oceania and Remote Oceania (Ingman and Gyllensten, 2003; Soares *et al.*, 2011). **B4a1a1a4** dates to ~4 ka and is found in the Bismarck Archipelago (Soares *et al.*, 2011). B4a1a1a2, not shown in the tree, carries nps 1473 and 3423A – the defining markers of the so-called “Malagasy motif”, identified in Madagascar by Razafindrazaka *et al.* (2010).

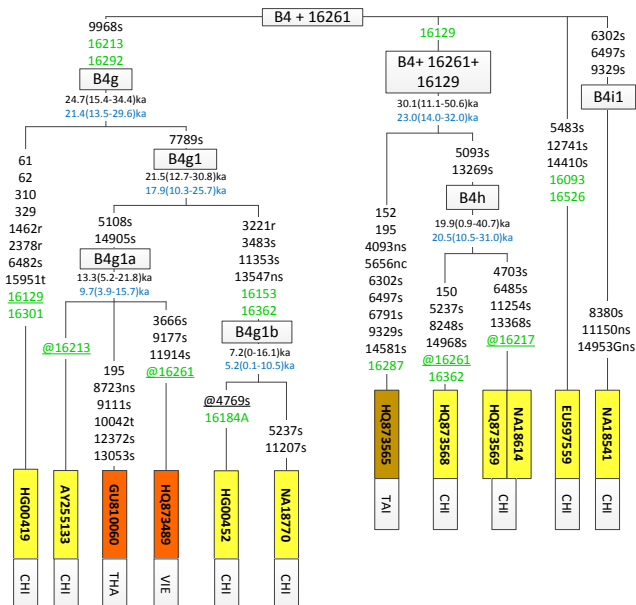
We can confirm this distribution with HVS-I data. Figure 5.8 shows the network for B4a1a1a, although it is labelled as B4a1 because it cannot be recognised with low-resolution HVS-I data. Furthermore, np 16247 is the control-region marker for B4a1a1a. Haplogroup B4a1a1a is common across Micronesia and Polynesia, but less common elsewhere. As seen in Figure 5.8, the majority of the root type is reported from coastal Papua New Guinea, where haplotypes from Eastern Indonesia (Moluccas and Nusa Tenggara) are much more diverse than those from the previous locations (Redd *et al.*, 1995). In ISEA, B4a1a1a is found in Alor, Ambon, Banjarmasin, Lombok, Manado, Toraja and Ujung Padang – located to the east of Southeastern Borneo and Lombok. The root type is seen in Sulawesi, but no derived types were found by Hill *et al.* (2007) hence suggesting a recent migration to Peninsular Malaysia from ISEA.

In Figure 5.9, **B4g** dates to ~25 ka with a single basal lineage seen in China (Zheng *et al.*, 2011). It then diverged into **B4g1**, dating to ~22 ka, which can be divided into **B4g1a**, dating to ~13 ka and seen in China (Kong *et al.*, 2003b), Thailand (Pradutkanchana, Ishida and Kimura, 2010) and Vietnam (Soares *et al.*, 2011), and **B4g1b**, dating to ~7 ka and seen in China (Zheng *et al.*, 2011). It is clearly recognisable by its HVS-I motif, and the HVS-I database confirms that it is found widely across southern China, Thailand and Vietnam.

B4+C16261T with a transition at np 16129 defines the **pre-B4h** node, it dates to ~30 ka and seen in the Siraya Taiwan (Soares *et al.*, 2011). It then diverged into **B4h**, which dates to ~20 ka, and seen only in China (Soares *et al.*, 2011; Zheng *et al.*, 2011).



**Figure 5.8 HVS-I network of B4a1 (it is in fact B4a1a1a). Figure adapted from Hill (2005).**



## 5.1.2 Haplogroups B4b'd'e'j and B4f

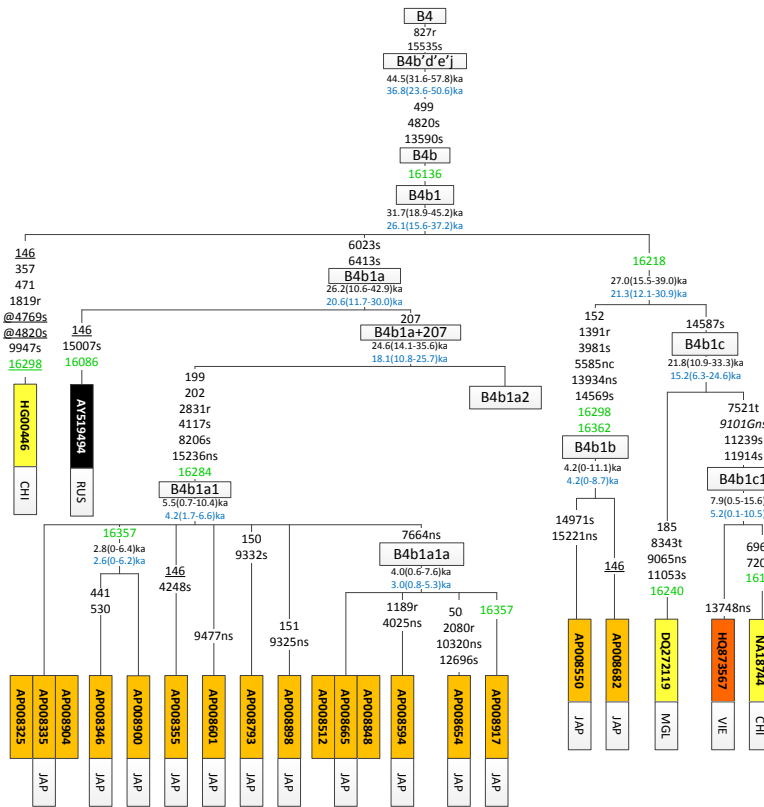


Figure 5.10 The tree of haplogroup B4b1 excluding B4b1a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, JAP – Japan, MGL – Inner Mongolia, China, RUS – Russia, VIE – Vietnam)

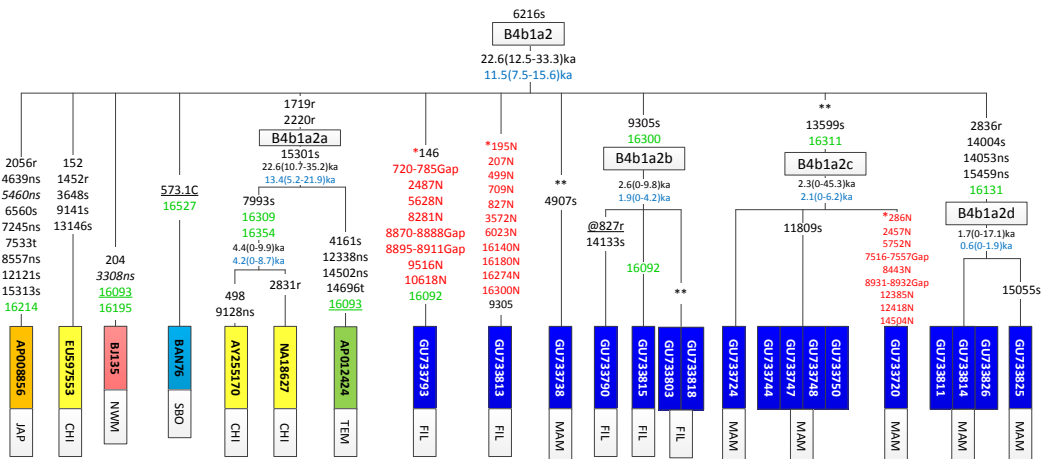


Figure 5.11 The tree of haplogroup B4b1a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. Sequences marked by “\*” are erroneous and not used in age calculation. Sequences with “\*\*\*” have np 310 removed as artefact since it forms incorrect evolutionary pathways in the clade. (CHI – China, FIL – Philippines, JAP – Japan, MAM – Philippines Mamanwa, NWM – Northwest Peninsular Malay, SBO – South Borneo, TEM – Aboriginal Malay Temuan)

**B4b'd'e'j** dates to ~45 ka and can be divided into B4b, B4d, B4e and B4j (Figure 5.10). Detailed descriptions for haplogroups B4d and B4f are available in Appendix E. B4b1 dates

to ~32 ka and the basal lineage is seen in south China (Zheng *et al.*, 2011). Two subclades nested within B4b1, B4b1a and B4b1+16218 (includes B4b1b and B4b1c). **B4b1a** dates to ~26 ka with a basal lineage seen in Siberia, Russia (Starikovskaya *et al.*, 2005). A further transition at np 207 dating to ~25 ka is the MRCA for B4b1a1 and B4b1a2. **B4b1a1** is entirely a Japanese clade (Tanaka *et al.*, 2004), dated to ~6 ka, and includes subclades **B4b1a1+16357** and **B4b1a1a**, both dated to ~3 ka and ~4 ka respectively.

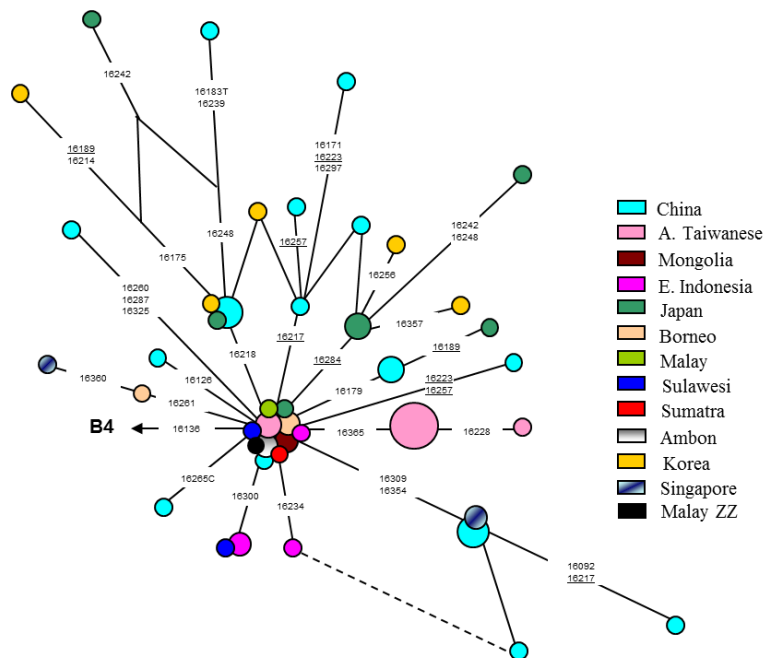
**B4b1a2** dates to ~23 ka (Figure 5.11) and is seen occasionally in Japan, China, Peninsular Malaysia and the Aboriginal Malay Temuan, but is most common in the negrito Mamanwa and the Philippines (six basal lineages). The ultimate source for B4b1 and B4b1a may be China, but the source for this subclade might therefore be Southeast Asia. B4b1a2 is divided into B4b1a2a, B4b1a2b, B4b1a2c, and B4b1a2d, of which the latter three are all restricted to the Philippines, and also include many paraphyletic lineages, found in the Malay Peninsula as well as the Philippines, with few in East Asia. **B4b1a2a** dates to ~23 ka where it is seen in the Temuan (Jinam *et al.*, 2012), before it splits into China, Guangdong (Kong *et al.*, 2003b) and Beijing (Zheng *et al.*, 2011) ~4 ka.

**B4b1a2b** dates to ~3 ka and seen in the Philippines Manobo and Surigaonon tribes (Gunnarsdóttir *et al.*, 2011a). Manobo is a non-negrito group and Surigaonon, an urban group in the Philippines. **B4b1a2c** dates to ~2 ka and found exclusively in the negrito Mamanwa (Gunnarsdóttir *et al.*, 2011a). **B4b1a2d** dates to ~2 ka, where it is seen only in the Surigaonons.

Haplogroup B4b (recognisable as B4b1 in HVS-I) is comparatively less common than B4a based on the HVS-I data. Figure 5.12 shows diverse haplogroup B4b types (derived and underived in HVS-I) are seen widely in SEA (in Ambon, Banjarmasin, Pekanbaru, Palu, Manado and Kota Kinabalu) and Taiwan, indicating a sampling lacuna in the whole-mtDNA data. One Malay individual is found with the root type and present at low levels similar to those from China, Korea, Mongolia, Taiwan Aborigines and one sample from Peninsular Malaysia (Figure 5.12). The derivative types can be seen in one from Banjarmasin with an additional mutation at np 16261 and an individual with a further transition at np 16380 from Singapore; another is seen in Palu with a transition at np 16300, which was also reported in three Eastern Indonesians by Redd *et al.* (1995) and is evident in the whole-mtDNA tree, represented by three samples from the Philippines. The pattern seems to suggest an ancient



East Asian source and more recent (but nevertheless possibly pre-Holocene) arrival in Taiwan, Peninsular Malaysia and ISEA.



**Figure 5.12 HVS-I network of B4b1. Figure adapted from Hill (2005).**

### 5.1.3 Haplogroup B4c

**B4c** dates to ~40 ka and can be divided into B4c1 and B4c2 (Figure 5.13). B4c is an entirely restricted to China and Japan except for B4c1b2a2, which is seen in SEA. Detailed descriptions for B4c1a and B4c1b are available in Appendix E.

**B4c1b2a2** dates to ~9 ka, and is widely distributed in SEA including the Philippines, Taiwan, Peninsular Malaysia, and Indonesia (Figure 5.14). At least four subclades nested within B4c1b2a2 that belong to the Philippines. The first subclade is defined by transitions at nps 3221, 12192, 13934 and 15734, which is seen only in the Manobos Philippines (Gunnarsdóttir *et al.*, 2011a). The second subclade is defined by a transition at np 4226 and dates to ~9 ka in Ivatan Philippines (Loo *et al.*, 2011), and it then further diverged into the Manobo (Gunnarsdóttir *et al.*, 2011a) ~1 ka. The third subclade dates to ~2 ka and seen in the Manobo and Surigaonon (Gunnarsdóttir *et al.*, 2011a). Lastly, the fourth subclade dates to ~9 ka and seen in the Ivatan Philippines (Loo *et al.*, 2011) and Sumatra (Gunnarsdóttir *et al.*,

2011b). B4c1b2a2 is also widely seen in Peninsular Malaysia. Two Minangkabau Malay formed a subcluster defined by transitions at nps 3666 and 15884, which dates to ~4 ka. The whole-mtDNA tree possibly suggests an origin of B4c1b2a in China at the beginning of sea-level rise ~15 ka, a time when Taiwan, Sumatra and Borneo might have been connected. The initial spreads might have taken along the Chinese and Sunda eastern coastlines towards Borneo and ultimately getting to the Philippines rather than via Taiwan.

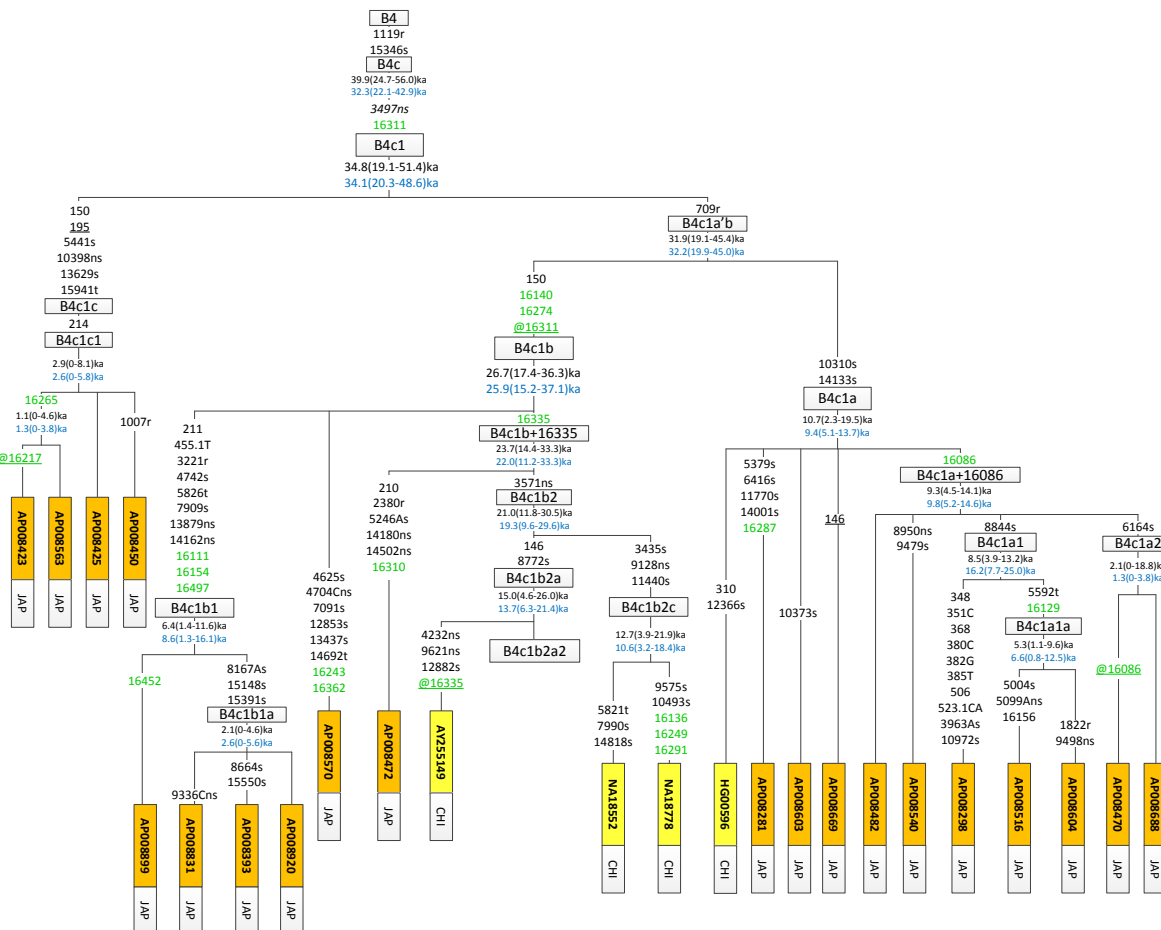


Figure 5.13 The tree of haplogroup B4c1 excluding B4c1b2a2. Time estimates shown for clades are ML and averaged distance ( $\rho$ ) in ka. (CHI – China, JAP – Japan)

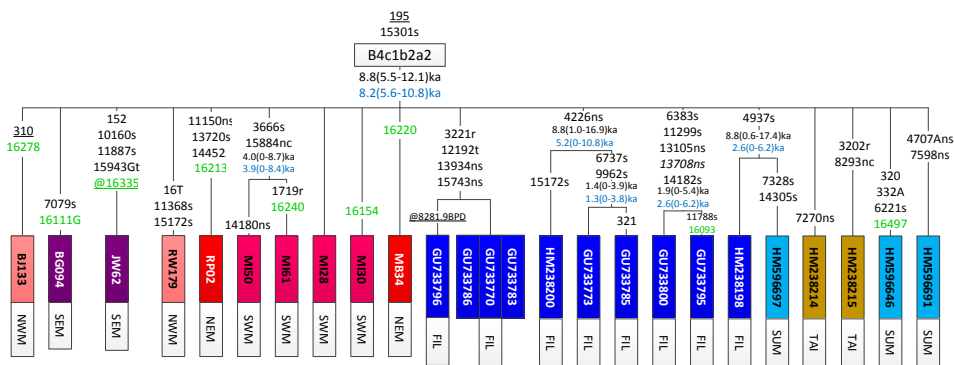


Figure 5.14 The tree of haplogroup B4c1b2a2. Time estimates shown for clades are ML and averaged distance (p) in ka. (FIL – Philippines, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, SWM – Southwest Peninsular Malay, TAI - Taiwan)

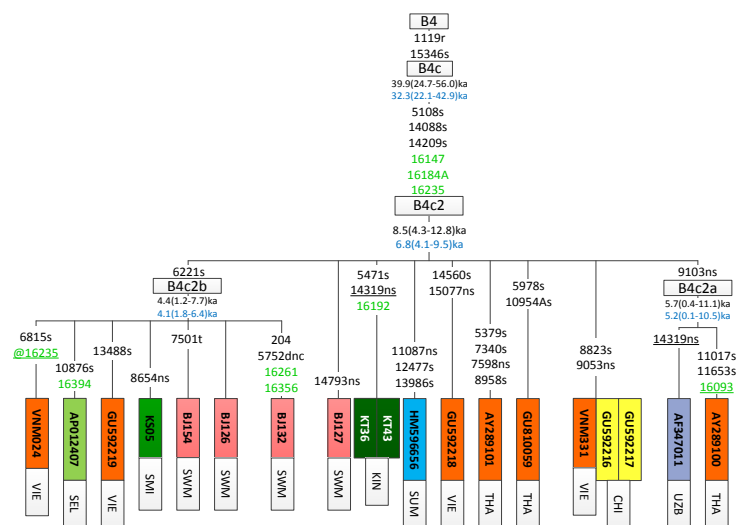


Figure 5.15 The tree of haplogroup B4c2. Time estimates shown for clades are ML and averaged distance (p) in ka. (CHI – China, KIN – Semang Kintak, SEL – Aboriginal Malay Seletar, SMI – Senoi Semai, SUM – Indonesia Sumatra, SWM – Southwest Peninsular Malay, THA – Thailand, UZB – Uzbekistan, VIE – Vietnam)

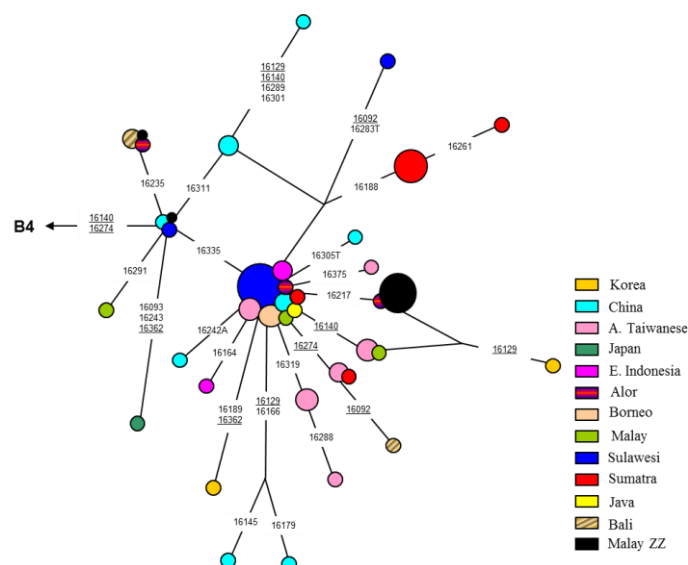


Figure 5.16 HVS-I network of B4c1b. Figure adapted from Hill (2005).

**B4c2** dates to ~9 ka and there are two subclades, B4c2a and B4c2b (Figure 5.15). B4c2 is seen in China (Peng *et al.*, 2010), Uzbekistan, Vietnam (Peng *et al.*, 2010; Archaeogenetics Research Group, Huddersfield), Thailand (Ingman and Gyllensten, 2003; Pradutkanchana, Ishida and Kimura, 2010), Peninsular Malaysia Malay and *Orang Asli* (the Semang Kintak and Aboriginal Malay Seletar), and Indonesia (Gunnarsdóttir *et al.*, 2011b). **B4c2a** dates to ~6 ka and seen in Uzbekistan (Ingman *et al.*, 2000) and Thailand (Ingman and Gyllensten, 2003). **B4c2b** dates to ~4 ka and it is seen in the Senoi Semai, Aboriginal Malay Seletar (Jinam *et al.*, 2012), Peninsular Malaysia, and Vietnam (Peng *et al.*, 2010; Archaeogenetics Research Group, Huddersfield).

Similar to haplogroup B4b, haplogroup B4c is also less common than B4a. Previous studies found B4c (recognisable as B4c1b in HVS-I) at relatively low levels in China, Taiwan, Peninsular Malaysia, Eastern Indonesia and Japan (Redd *et al.*, 1995; Seo *et al.*, 1998; Kivisild *et al.*, 2002; Yao *et al.*, 2002a; Tajima *et al.*, 2003; Zainuddin and Goodwin, 2004). Among the ISEA samples in Hill (2005), haplogroup B4b1c was found in 9 individuals from Sulawesi (two Manado, three Ujung Padang and four Toraja), Pekanbaru, Sumatra, and lesser in Alor, Bali, Borneo and Sumatra. Figure 5.16 shows the root type of B4c1b was found in China and Sulawesi, where modern Malay also found to have the same type. The one-step derivatives were seen in individuals from Alor and Bali with transition at np 16235, one Malay individual with a further np 16291, and three from China at np 16311. 21 Malay individuals are found with haplogroup B4c1\* which is further defined by a transition at np 16335, previously found in one Alor by Hill (2005). Now it is evident from the whole-mtDNA tree that this represents subclade B4c1b2a2, which possibly suggests it has a source in SEA during early Holocene ~9 ka.

## 5.2 Haplogroup B5

The other major branch of B is haplogroup B5, dating to ~52 ka. The B5 tree here includes 82 complete sequences, equally representing the two subclades: 41 B5a and 41 B5b. Figure 5.17 shows the major further subclades of haplogroup B5. Like B4, B5 seems to have a southern origin in MSEA/southern China and then spread into SEA, shows particularly in subclades B5a1a, B5a1b, B5a1c and B5b1c.

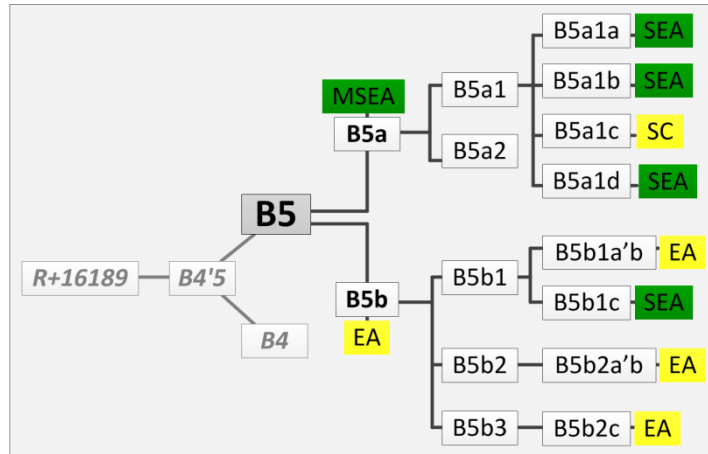


Figure 5.17 Schematic diagram of haplogroup B5 and its major subclades distribution. (EA – East Asia, MSEA – Mainland Southeast Asia, SC – Southern China, SEA – Southeast Asia)

### 5.2.1 Haplogroup B5a

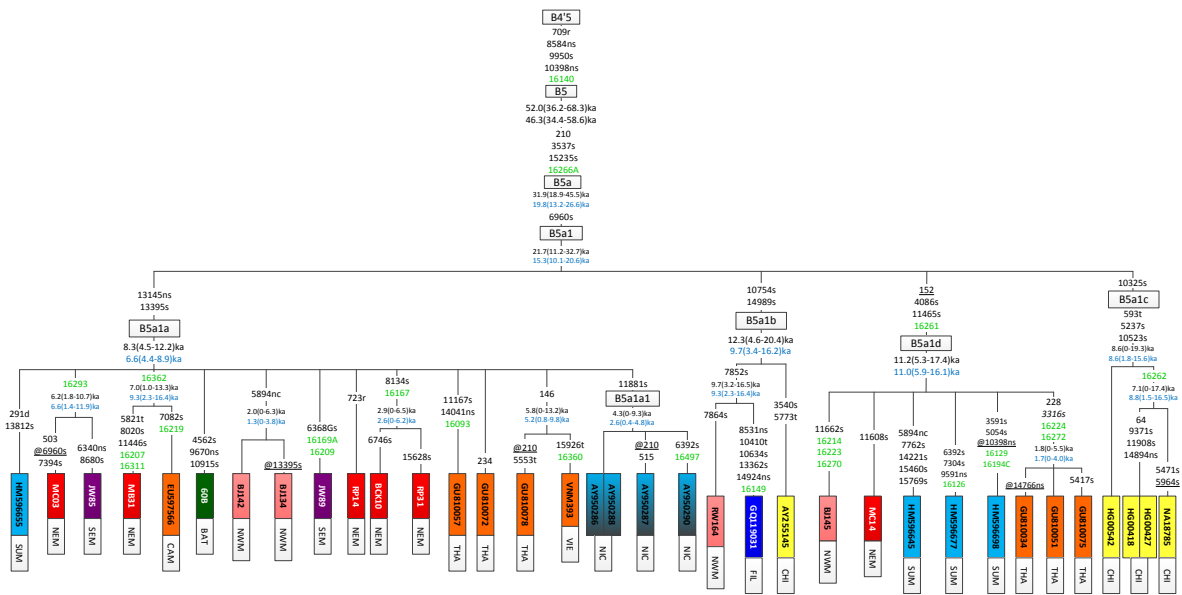
**B5a** dates to ~32 ka and it has two subclades, B5a1 and B5a2. **B5a1** dates to the LGM ~22 ka and is divided into B5a1a, B5a1b, B5a1c, and B5a1d (Figure 5.18). **B5a1a** dates to ~8 ka and it is widely seen in SEA, including Peninsular Malaysia, the Semang Batek (and one Senoi Temiar from HVS-I), Cambodia (Hartmann *et al.*, 2009), Indonesia Sumatra (Gunnarsdóttir *et al.*, 2011b), Thailand (Pradutkanchana, Ishida and Kimura, 2010), Vietnam and the Nicobar Islands (Thangaraj *et al.*, 2005). B5a1a includes B5a1a1 and five other unnamed subclades. B5a1a1 is exclusively seen in the Austro-Asiatic-speaking inhabitants of the Nicobar Islands (Thangaraj *et al.*, 2005) and dates to ~4 ka. Other documented Austro-Asiatic speakers nested within B5a1a including the so-called negrito Semang Batek (who speaks Northern Aslian) and Senoi Temiar (Central Aslian-speaker, HVS-I data).

Subclade B5a1a with a further transition at 16293 dates to ~6 ka and is seen in Northeast and Southeast Peninsular Malaysia (this study). Subclade B5a1a with a further transition at np 16362 dates to ~7 ka and it is seen in Northeast Peninsular Malaysia (this study) and Cambodia (Hartmann *et al.*, 2009). Subclade B5a1a with a transition at np 5894 is seen in Northwest Peninsular Malaysia (this study) and dates to ~2 ka. Subclade B5a1a with two transitions at nps 8134 and 16167 dates to ~3 ka and is seen in Northeast Peninsular Malaysia (this study). Lastly, subclade B5a1a with a transition at np 146 dates to ~6 ka and this subclade is seen in Vietnam (Archaeogenetics Research Group, Huddersfield) and Thailand (Pradutkanchana, Ishida and Kimura, 2010). It is clear that B5a1a and its subclades are restricted to MSEA, Peninsular Malaysia and Nicobars Islands. Bellwood (1997) suggested that the Austro-Asiatic-speaking foragers have an origin in South China during

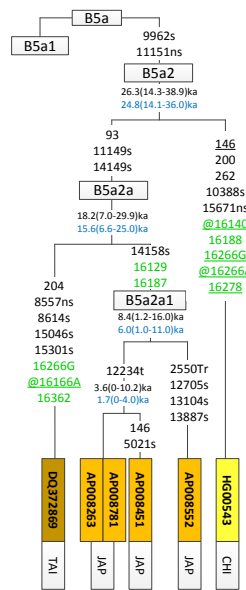
middle Holocene. The whole-mtDNA tree appears to be consistent to Bellwood's view that the Austro-Asiatic-speakers can trace their origin to South China/Mainland SEA during middle Holocene ~8 ka. Bellwood (2001) also pointed out that one of the characteristics that indicates the spread of a linguistic family together with a Neolithic farming expansion is the rapid spread of a language over a large area. The fact that B5a1a and its subclades dispersed across MSEA, Peninsular Malaysia and Nicobar Islands between 2 ka to 7 ka correspond to the coastal Neolithic expansions suggested by Bulbeck (2008), perhaps in several events (also Higham, 2004). Although in this case, the lineages from Peninsular Malaysia and Sumatra, Indonesia would be the Austronesian-speaking farmers, who have arrived in Sumatra from MSEA and Peninsular Malaysia and not via the "Out of Taiwan" route.

**B5a1b** dates to ~12 ka, where the basal lineage is seen in Hubei China (Kong *et al.*, 2003b). It then dispersed ~10 ka into the Philippines (Tabbada *et al.*, 2010) and Northwest Peninsular Malaysia. **B5a1d** dates to ~11 ka, it is seen in northern Peninsular Malaysia (this study), Sumatra, Indonesia (Gunnarsdóttir *et al.*, 2011b) and Thailand (Pradutkanchana, Ishida and Kimura, 2010), the latter forms a subclade which dates to ~2 ka. **B5a1c** dates to ~9 ka, and a subclade nested within dates to ~7 ka. **B5a1c** is restricted to south China only (Zheng *et al.*, 2011). In general, B5a1 shows a post-glacial Sunda distribution with recent offshoots. The small subclade of B5a1b is likely to show the relict descendants that survived since the early Holocene dispersal and found in Hubei China, the Philippines and Peninsular Malaysia.

**B5a2** dates to pre-LGM ~26 ka and the basal lineage is seen in south China (Figure 5.19; Zheng *et al.*, 2011). It is further divided into **B5a2a**, dating to ~18 ka, nesting an aboriginal Taiwanese lineage (Pierson *et al.*, 2006), which in turn nests a subclade seen only in Japan (Tanaka *et al.*, 2004), dating to ~8 ka – a pattern suggesting a dispersal from Southeast Asia into Northeast Asia by the early Holocene.



**Figure 5.18** The tree of haplogroup B5a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BAT – Semang Batek, CAM – Cambodia, CHI – China, FIL – Philippines, JAP – Japan, NEM – Northeast Peninsular Malay, NIC – Nicobars, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, TAI – Taiwan, THA – Thailand, VIE - Vietnam)



**Figure 5.19** The tree of haplogroup B5a2. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, TAI – Taiwan)

## 5.2.2 Haplogroup B5b

**B5b** dates to ~33 ka and divided into B5b1, B5b2 and B5b3 (Figure 5.20). **B5b1** dates to ~28 ka and single basal lineages are seen in Japan (Tanaka *et al.*, 2004) and China (Zheng *et al.*, 2011), likely to have an origin in China. B5b1 has two subclades, B5b1a'b and B5b1c.

**B5b1a'b**, seen only in Japan, dating to ~22 ka and divided into B5b1a and B5b1b. **B5b1a** dates to ~8 ka and is found in Japan Aichi and Tokyo (Tanaka *et al.*, 2004). **B5b1b** dates to ~5 ka, and nested within a subclade defined by a transition at np 14959 dates to ~2 ka, and found in Tokyo, Chiba and Aichi.

**B5b1c** has a very distinctive distribution in the context of B5b. There are basal lineages in the Semang Batek, Peninsular Malaysia, and the Philippines (Tanaka *et al.*, 2004; Gunnarsdóttir *et al.*, 2011a; Loo *et al.*, 2011), dating to ~11 ka. A potential subclade nested within B5b1c is seen in the Philippine negrito Mamanwa clustering with the Manobos (Gunnarsdóttir *et al.*, 2011a); although these sequences were not used in age estimations since there were many gaps. These lineages possibly are the relict descendants that survived in the Sunda after the second flood, about 11 ka, and incidentally linking the Austro-Asiatic-speakers Semang Batek and the Austronesian-speakers Philippines Mamanwa within the same clade.

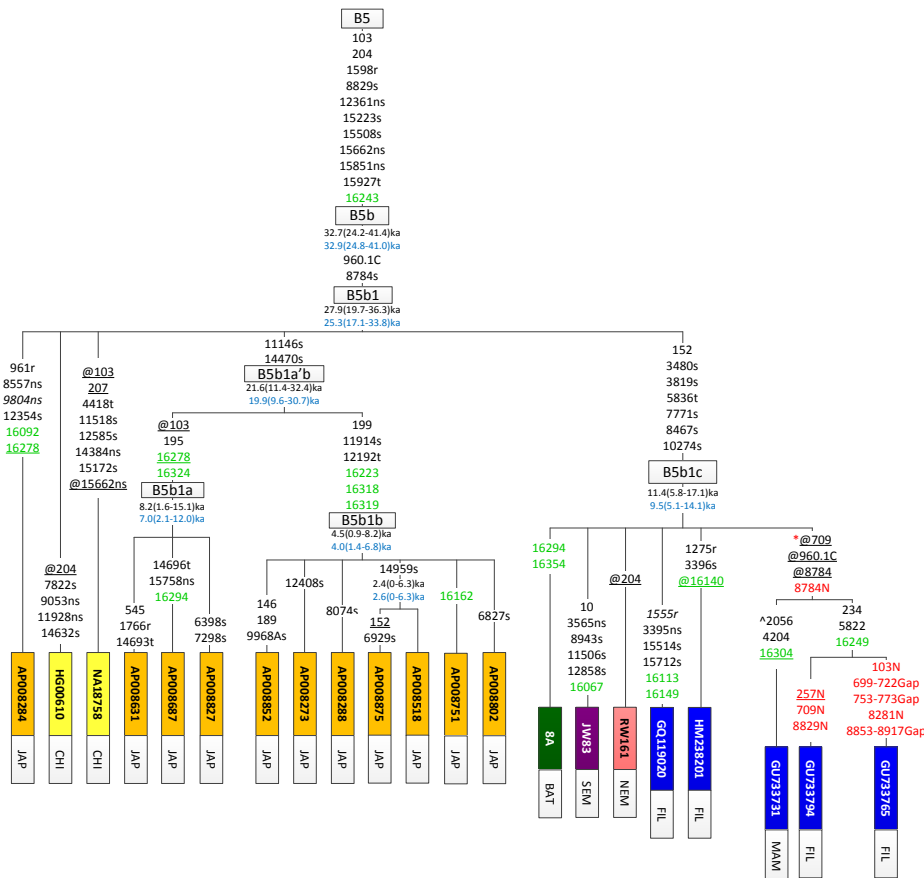
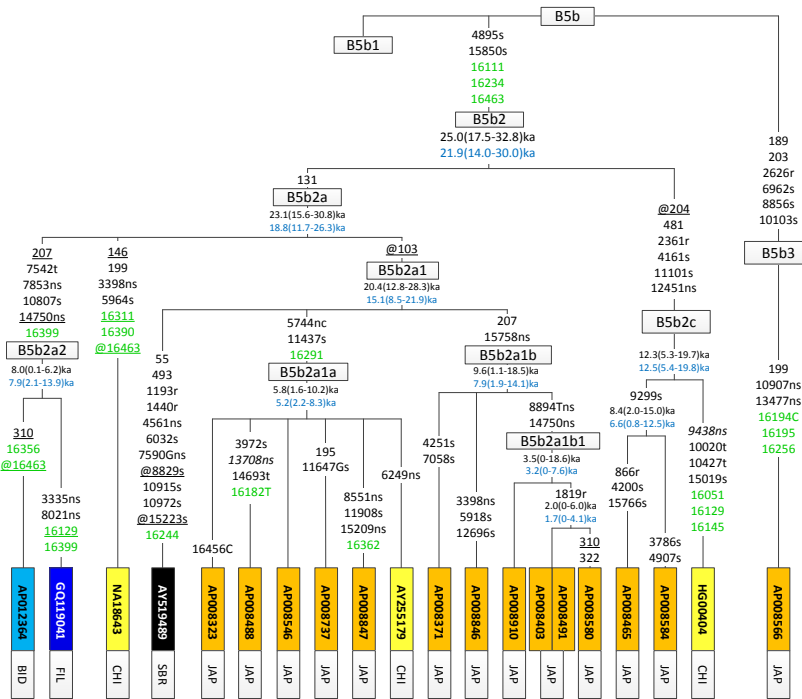
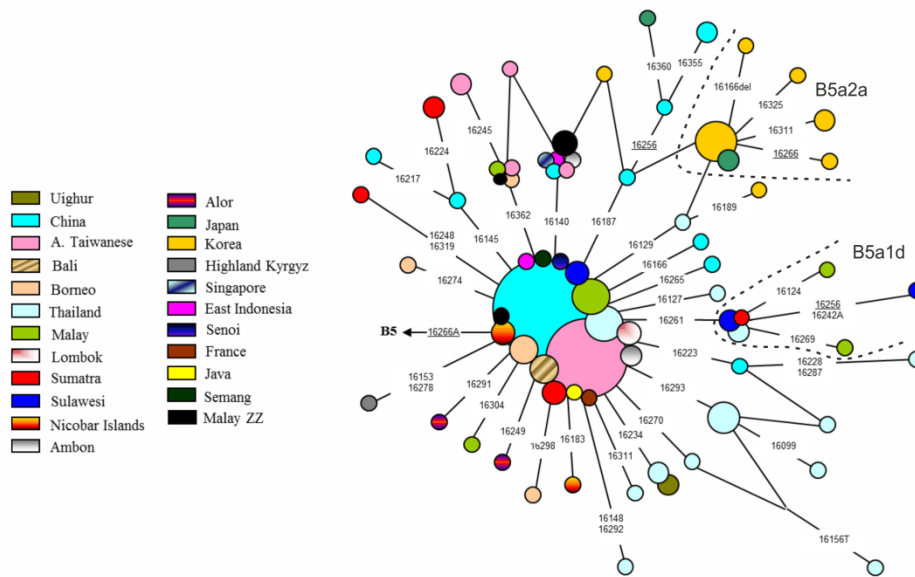


Figure 5.20 The tree of haplogroup B5b1. Time estimates shown for the clades are ML (in black) and averaged distance (p; in blue) in ka. (BAT – Semang Batek, CHI – China, FIL – Philippines, JAP – Japan, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, SEM – Southeast Peninsular Malay)





**Figure 5.21** The tree of haplogroup B5b2 and B5b3. Time estimates shown for the clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. (BID – Sarawak Bidayuh, CHI – China, FIL – Philippines, JAP – Japan, SBR- Siberia, Russia)



**Figure 5.22** HVS-I network of B5a. Figure adapted from Hill (2005).

**B5b2** dates to ~25 ka and consists of two subclades, B5b2a and B5b2c, distributed mainly across China and Japan (Figure 5.21). **B5b2a** dates to ~23 ka, with a basal lineage seen in Beijing, China (Zheng *et al.*, 2011). B5b2a has two subclades, B5b2a1 and B5b2a2 (newly named here). **B5a2a1** dates to ~20 ka, with a single basal lineage in Russian Siberia (Starikovskaya *et al.*, 2005), and two subclades: **B5b2a1a** and **B5b2a1b**, dating to ~6 ka and

~10 ka respectively, seen only in Japan (Tanaka *et al.*, 2004) and China (Kong *et al.*, 2003b). B5b2a1 looks to have an origin in China, dispersing into Japan in roughly the mid-Holocene.

**B5b2c** dates to ~12 ka, and shows a roughly similar pattern to B5b2a1, with a basal lineage in south China (Zheng *et al.*, 2011), and the main subclade seen in Japan (Tanaka *et al.*, 2004) around 8 ka. **B5b3** is represented here by a sample from Japan (Tanaka *et al.*, 2004).

Figure 5.22 shows the HVS-I data network of B5a adapted from Hill (2005). It is very poorly resolved in comparison with the whole-mtDNA tree, as there are few informative HVS-I sites within the tree. B5a1d and B5a2a are recognisable and outlined in the network. Apart from those already present in the whole-mtDNA B5a1 tree (such as MSEA, Peninsular Malaysia and Nicobar Islands), other potential B5a1 lineages might include the Aboriginal Taiwanese, Sulawesi, Java, Moluccas of Indonesia, and Singapore. The network also shows that B5a1d may expand as far east as Sulawesi, and B5a2a may not just be restricted to Japan but the root type was also found in Korea.

Similar to most HVS-I network, network of B5b has low resolution and lack of informative HVS-I sites compared with the whole-mtDNA tree. In previous studies, haplogroup B5b was commonly found in China, Japan, Korea, the Philippines and Micronesia (Lee *et al.*, 1997; Lum *et al.*, 1998; Seo *et al.*, 1998; Yao *et al.*, 2002a, 2002b). We now know from the whole-mtDNA tree that B5b1a'b and subclades are restricted to Japan, and B5b1c are seen only in Peninsular Malaysia and the Philippines. The possible B5b1c lineages recognisable from the HVS-I network include Sulawesi, Borneo, East Indonesia, Sumatra, Lombok, and Singapore, where the "Malay ZZ" and Semang Batek are represented in the whole-mtDNA tree. Haplogroups B5b2 and B5b2a1a are recognisable in the HVS-I network and outlined in Figure 5.23. The network is consistent with the results shown in the whole-mtDNA tree, where B5b2 has an East Asia origin, where plenty of derived sequences are found in China, Korea, Japan, and one each from Inner Mongolia and Kyrgyz. The Sumatran lineage might possibly correspond to the lineages from the Philippines and Bidayuh, North Borneo in the whole-mtDNA of subclade B5b2a2. The much better resolution afforded by the whole-mtDNA tree shows that the Southeast Asia clades are shallower, so that the suggestion in the HVS-I network that a Southeast Asian origin might be possible is not supported by the new analysis. The source most likely lies in China with greatest diversity and with post-glacial dispersals both ways.

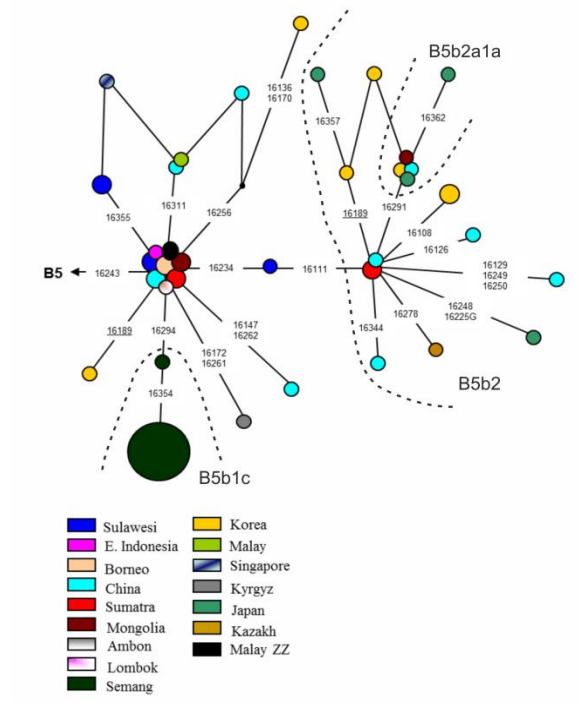


Figure 5.23 HVS-I network of B5b. Figure adapted from Hill (2005).

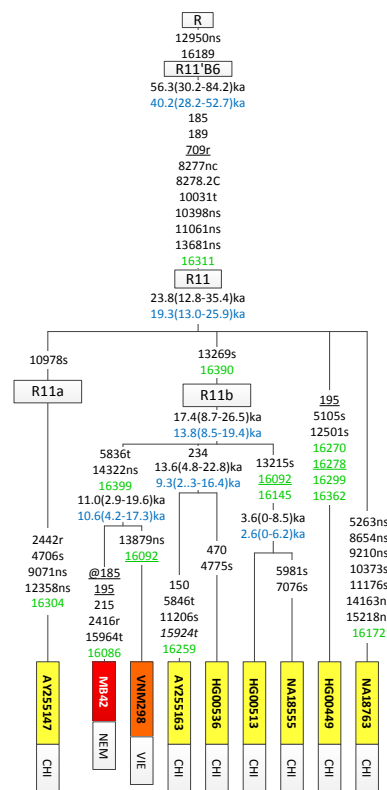


Figure 5.24 The tree of haplogroup R11. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, NEM – Northeast Peninsular Malay, VIE – Vietnam)



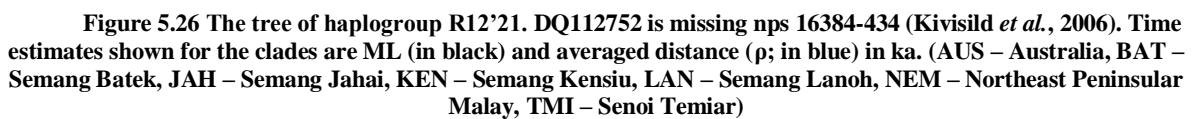
**B6**, on the other hand, is found predominantly in MSEA/Malay Peninsula and dates to ~27 ka (Figure 5.25). It is divided into B6a and B6b (new nomenclature). A basal type of B6 is seen in Vietnam (Archaeogenetics Research Group, Huddersfield). **B6a** dates to ~18 ka, with a basal lineage in the Philippines (Tabbada *et al.*, 2010) and a subclade, **B6a1** dating to ~10 ka with a basal lineage in Thailand. It further diverges into a subclade, **B6a1a**, which appears restricted to the Malay Peninsula, dating to ~7 ka, and shared by several Malay and Temuan Aboriginal Malay (one of them is reported by Jinam *et al.*, 2012).

**B6b** dates to ~23 ka, and even rarer, seen in one individual from China (Kong *et al.*, 2003b) and one from Vietnam (Archaeogenetics Research Group, Huddersfield).

## 5.4 Haplogroup R12'21

**R12'21** is basal to haplogroup R. The ancestral node is inadequately dated to ~72 ka by  $\rho$  since the mtDNA sequence of **R12** Aboriginal Australian sample is an incomplete sequence with a gap between np 16384 and np 434 (Kivisild *et al.*, 2006; Hudjashov *et al.*, 2007). The R12 sequence is also excluded from ML calculation for this reason. The R21 tree includes 14 complete sequences. The unique deep phylogenetic link to Australia, however, supports the great local antiquity of this lineage on the Sunda shelf in SEA, whatever the uncertainty on the  $\rho$  age estimate (Figure 5.26).

**R21** dates to ~12 ka, where the root type splits between the *Orang Asli* and Malay in Kelantan Malaysia. The young divergence time of R21 is due to genetic drift and population subdivision of the *Orang Asli*. They form a clade, **R21a**, which dates to only ~6 ka. It is concentrated in the northern Semang, especially the Jahai (Jinam *et al.*, 2012) and Senoi Temiar, the latter borders the Jahai in Kelantan. Jahai are reported to be in frequent contact with speakers of Malay, and many share settlements with speakers of Temiar (a Central Aslian language) (Burenhult, 2001). The deep diversity preserved in the Malay shows that although it has undergone drift, the ancient diversity captured in the large Malay population indicates a markedly greater diversity than in the small relict *Orang Asli* populations.



Haplogroup **R22** is basal within R and dates to ~46 ka (Figure 5.27). R22 has three distinctive subclades, which are hereby nominated as R22a, R22b and R22c. R22 is here reconstructed from eight complete sequences. Haplogroup R22a and R22b appeared to have experienced high drift.

185



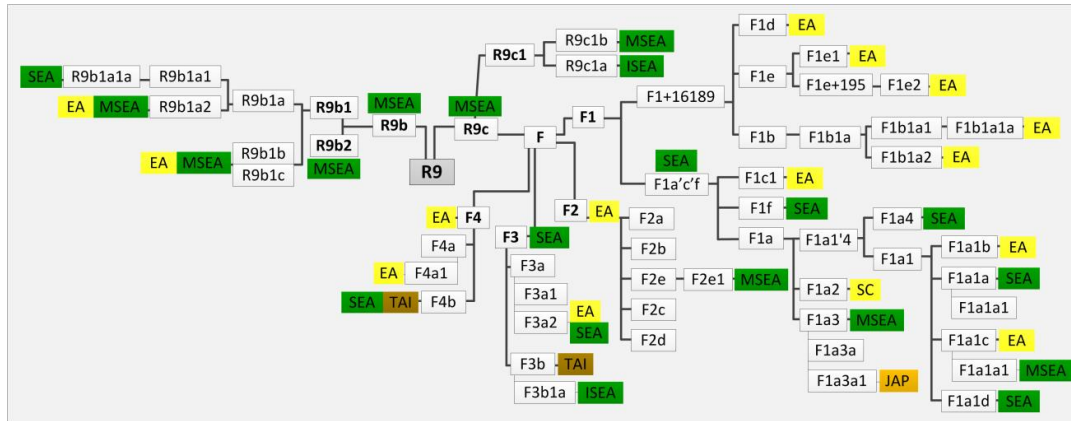
Islands (Prasad *et al.*, 2001) and Thailand (Yao *et al.*, 2002b). HVS-I data (Figure 5.28) showed that R22a includes individuals from Bali, Borneo, Sumba, Lombok and Thailand. An HVS-I subclade defined by a reversion at np 16304, which is likely to be R22b, is seen in Sumatra, Sumba, Sulawesi, and Lombok. The HVS-I signature of the Nicobar Islanders simply form a new R22 subclade.

R22 appears to be a surviving lineage from the initial founding of the Sunda shelf region and is widespread and present in mainstream indigenous groups. The origin of R22, given the age and distribution of the analysis, would have been Sundaland.

## 5.6 Haplogroup R9

R9 is defined by polymorphisms at nps 3970, 13928C and 16304, with two main branches: R9b and R9c (which includes F) (Kong *et al.*, 2003b; Soares *et al.*, 2009). These two diverged from the R9 root ~55 ka. The phylogeny of R9 here includes 201 complete sequences: 24 R9b, 8 R9c1 and 168 F. R9 appears, from the pattern shown in the schematic diagram below (Figure 5.29), to have originated in MSEA and spread both ways throughout China and SEA. **R9b** has deep roots in MSEA and the sink recipients for R9b1a, R9b1b and R9b1c are seen in ISEA in the south and East Asia in the north. R9c is further divided into R9c and F. **R9c1** is a rare haplogroup that has two subclades, R9c1a and R9c1b. R9c1a has dispersed into ISEA, while R9c1b is seen in China and Vietnam, again indicating an origin in MSEA. Haplogroup **F** has four subclades, F1, F2, F3, and F4. **F1** is split into F1+16189 and F1a'c'f. F1+16189 is common in China and Japan. On the other hand, F1a'c'f appears to have deep roots in SEA, where the subclades are widely distributed throughout Island and Mainland SEA. F1f potentially has a source in western Sunda and dispersed into ISEA. However, several subclades like F1c1, F1a1b, F1a1c, and F1a3a1, appear to settle recently in China and Japan; while F1a2 in South China. **F2** suggests a root in China and Japan for its five subclades, except for F2e1 that is seen in MSEA. **F3** is widely distributed throughout East Asia and SEA. **F4** is seen in China and Japan, where the sink recipients of F4b are almost entirely restricted to Taiwan and also seen in SEA.





### 5.6.1 Haplogroup R9b

Subclade R9b is seen at low levels in MSEA: West Malaysia (Malay and Aboriginal Malays, both Semelai and Temuan; Hill *et al.*, 2006), Vietnam, Thailand, and also in Indonesia (Hill *et al.*, 2007). R9b has also been found at low rates in the Yunnan and Guangxi provinces of South China (Yao and Zhang, 2002). R9b dates to ~46 ka, and is divided into R9b1 and R9b2. **R9b1** dates to ~25 ka and it has three subclades: R9b1a, R9b1b and R9b1c (Figure 5.30). The deepest lineages of R9b1 are found both in South China (Kong *et al.*, 2003b) and MSEA: Vietnam (Hill *et al.*, 2006). Similarly, **R9b1b** and **R9b1c**, dated to ~4 ka and ~11 ka respectively, are both found only in China (Zheng *et al.*, 2011) and Vietnam (Hill *et al.*, 2006).

**R9b1a** dates to the LGM ~20 ka, and includes two nested subclades, **R9b1a1** and **R9b1a2** (Figure 5.30). The basal branch of **R9b1a1** is seen in the Philippine negrito Mamanwa (Gunnarsdóttir *et al.*, 2011a), and an additional nested subclade, **R9b1a1a**, is found only in MSEA and the Greater Sundas, dating to ~10 ka. **R9b1a2** dates to ~6 ka, and is found both in China (Kong *et al.*, 2003b) and Vietnam (Hill *et al.*, 2006). **R9b1a1a** has three subclades, found only in SEA: **R9b1a1a1** dates to ~7 ka and is seen only in Indonesia (Sumatra, Java and Sulawesi); **R9b1a1a2** dates to ~8 ka and, being confined to MSEA, presumably took the Peninsular route from Thailand (Hill *et al.*, 2006) into the Aboriginal Malays (Semelai) ~1 ka; **R9b1a1a3** presumably also spread south down the Malay Peninsula ~10ka, where it is now found among Northeast Peninsular Malay, as well as forming a subclade aged around ~5 ka among the Semang Kintak of the north-western interior and Aboriginal Malays further south, the former forming a further subclade dating to ~1 ka.

**Haplogroup R9b2** dates to ~6 ka, and, again, is found only in MSEA: Thailand, Vietnam (Hill *et al.*, 2006; Pradutkanchana, Ishida and Kimura, 2010) and Northeast Peninsular Malay.

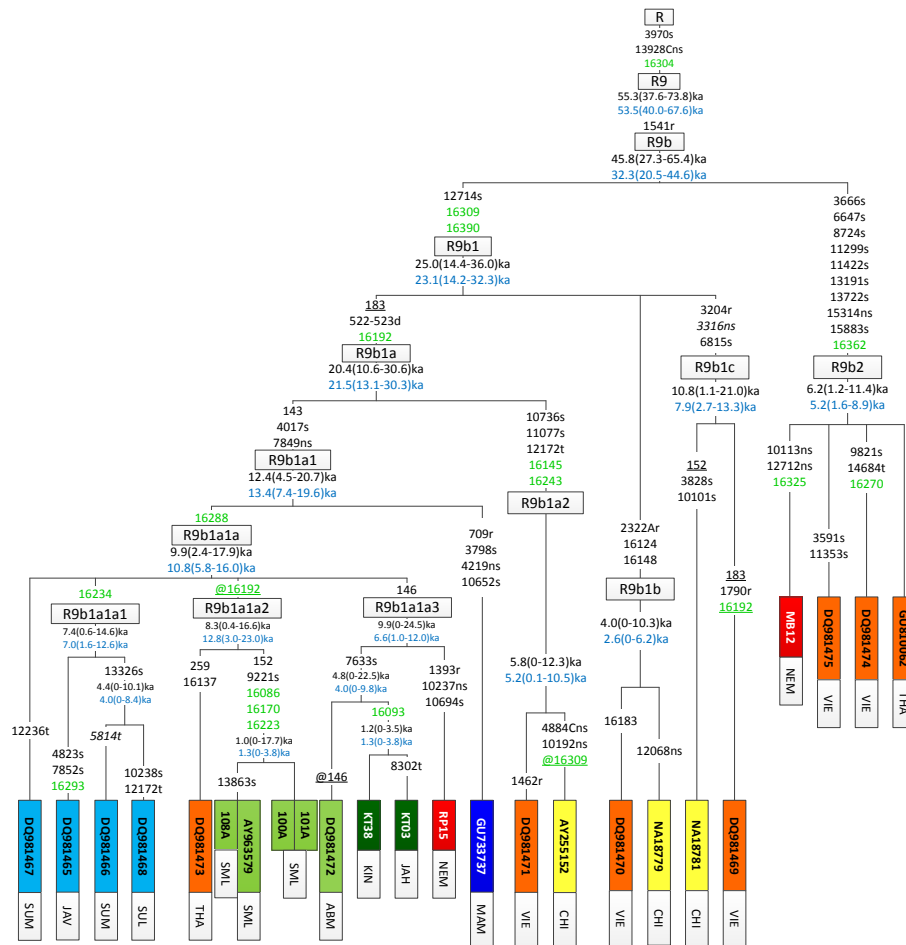


Figure 5.30 The tree of haplogroup R9b. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (ABM – Aboriginal Malay, JAH – Semang Jahai, JAV – Java, Indonesia, KIN – Semang Kintak, MAM – Philippines Mamanwa, NEM – Northeast Peninsular Malay, SML – Aboriginal Malay Semelai, SUL – Sulawesi, SUM – Sumatra, THA – Thailand, VIE – Vietnam)

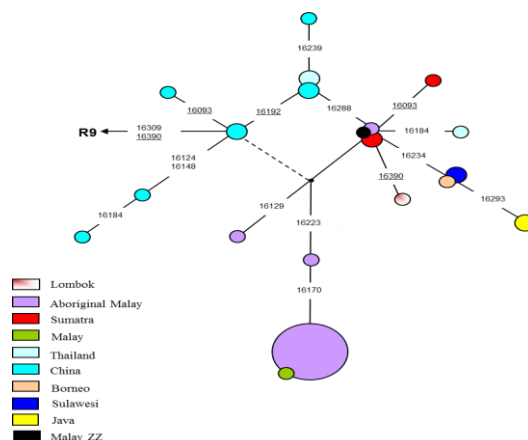


Figure 5.31 HVS-I network of R9b1 types. Figure adapted from Hill (2005).

Figure 5.31 shows the HVS-I network of haplogroup R9b (R9b1 in current whole-mtDNA nomenclature). In this early representation, the more basal types with the transition at nps 16192 (recognisable as R9b1a in the HVS-I network) are found in Thailand and South China (Yunnan, Guangxi; Yao *et al.*, 2002b) and Xinjiang of North China (Yao *et al.*, 2002a). The types with a further transition at np 16288 (currently R9b1a1a in the whole-mtDNA tree) are found in Borneo, Lombok, Sulawesi and Java. The most common type in this early survey was represented largely by one sequence type in the Aboriginal Malays (Semelai and Temuan) (Hill *et al.*, 2007) and one “Malay ZZ” (identifiable now as R9b1a1a2 in HVS-I network). The dotted line in Figure 5.31 showed that the HVS-I network was not able to clarify the polarity of evolution at np 16192 (Hill, 2005; Hill *et al.*, 2006), but with the new whole-mtDNA data presented here, it confirms that the Aboriginal Malays and “Malay ZZ” (and a Thai lineage in Figure 5.30) has loss np 16192. Although the HVS-I network seems to point to an origin in China, the whole-mtDNA tree indicates that the source was most likely Sundaland, given the age and taking into account the distribution of R9b2 in Vietnam and the Malay.

### 5.6.2 Haplogroup R9c1

**R9c1** dates to ~33 ka and includes two small subclades (Figure 5.32). R9c1a dates to ~8 ka; it is recognisable from a clear diagnostic HVS-I motif and is seen most frequently in Taiwan, and less commonly in the Philippines and Indonesia. Two of the Philippine negrito Batak (Scholes *et al.*, 2011) lineages form a nested subclade that dates to ~6 ka. The Batak, one of the Philippine negrito groups found on Palawan Island, speak Austronesian languages instead of the non-Austronesian languages thought to have been spoken before the Holocene (Reid, 1994; Gray *et al.*, 2009) and lead a hunter-gatherer lifestyle. Their geographically interesting location offers the possibility of movement via near-land-bridges between Sabah in East Borneo and Palawan in the Philippine Archipelago, in regards to historical population interactions in the region (Scholes *et al.*, 2011), though the presence of a Taiwanese haplotype in R9c1a could also suggest a north-south (or south-north) maritime movement. The Philippine Batak appears to have common ancestry with the non-negrito on Palawan as well as the neighbouring regions, which is observed in R9c1a.

**R9c1b** (Figure 5.32), dating to ~15 ka, is seen in only three individuals from South China and Vietnam (Zheng *et al.*, 2011; Archaeogenetics Research Group, Huddersfield); in the HVS-I database it is recognisable by default as R9c1 lacking the full R9c1a motif and is

seen across southern China/MSEA, where it likely originated, with occasional individuals in ISEA and Taiwan.

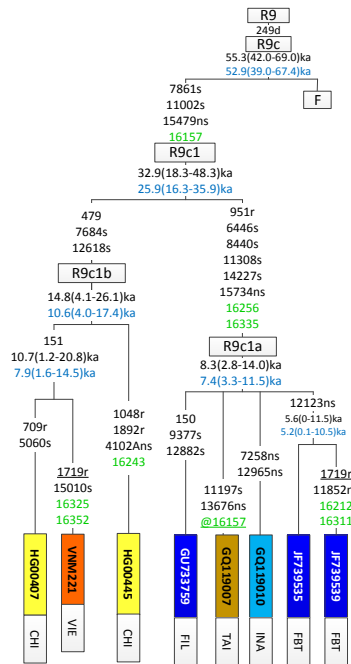


Figure 5.32 The tree of haplogroup R9c1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, INA – Indonesia, FBT – Philippines Batak, FIL – Philippines, TAI – Taiwan)

### 5.6.3 Haplogroup F1

Haplogroup F is the main subclade of R9c and dates to ~51 ka, dividing into F1, F2, F3 and F4, with no basal lineages seen to date, either for F or for its major subclade, F1. The main subclade of F, haplogroup **F1**, dating to ~38 ka, and has two subclades called F1+16189 (the hypothetical immediate precursor of F1b, F1d, F1e and F1g) and F1a’c’f. F1+16189, if it represented a true clade, would date to ~35 ka and with an origin in China and Japan. For detailed descriptions of haplogroups F1b, F1d and F1e, all not seen in Malaysia, see Appendix E.

In Figure 5.33, **F1a’c’f** dates to ~29 ka. The clade is represented extensively both in China and SEA. It includes F1c and F1a’c’f+16172 (including F1a and F1f). From the complete mtDNA tree, with only a few sequences, it appears that F1c and F1f have quite different distributions, but as they are not distinguishable in HVS-I it is impossible to confirm or augment this observation with more data at present. **F1c1**, the sole branch of F1c, dating to ~19 ka with basal branches seen in Japan (Tanaka *et al.*, 2004) and China (Kong *et al.*, 2003b; Zheng *et al.*, 2011). F1a’f dates to ~25 ka. **F1f** dates to ~6 ka and is seen in Malay

from across the Peninsula (including a direct match to several Bidayuh from Borneo, at the root of F1f), the Sarawak Bidayuh (Jinam *et al.*, 2012) and Sumatrans (Gunnarsdóttir *et al.*, 2011b), with a single individual from Beijing, China (Zheng *et al.*, 2011). Several Sumatrans form a subclade that dates to ~2.5 ka, with a further transition at np 8490 dating to ~1 ka. This shows a close connection between Peninsular Malay and likely source populations in Sumatra and, especially Sarawak in Borneo, who all speak the Austronesian Malayo-Polynesian languages. The results here therefore show that F1f is likely to have a source in west Sunda with dispersal into ISEA ~6 ka.

**Figure 5.33 The tree of haplogroup F1a'c'f excluding F1a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\bar{p}$ ; in blue) in ka. (\*\* – np 310 removed; BID – Bidayuh Sarawak, CHI – China, FBT – Philippines Batak, FIL – Philippines, JAP – Japan, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, SUM – Sumatra, SWM – Southwest Peninsular Malay)**

and a Filipino (Family Tree DNA, 2006; Loo *et al.*, 2011). A small subclade nested below named **F1a3a1**, dating to ~6 ka, and so far is only seen in Japan (Tanaka *et al.*, 2004). F1a3 was considered a possible marker for Austronesian dispersal by Hill *et al.* (2007); its age and distribution now make this less likely.

**F1a1'4** dates to ~20 ka and splits into F1a4a and F1a1, dating to ~7 ka and 15 ka respectively. The HVS-I database shows that **F1a4a** is dispersed thinly but widely in South China, Taiwan aboriginals and both MSEA and ISEA, with examples in the complete-mtDNA tree from the Northeast Peninsular Malay, Sumatra and the Philippines (Gunnarsdóttir *et al.*, 2011a; Loo *et al.*, 2011). F1a4 was proposed as a possible marker for the Austronesian dispersal in Hill *et al.* (2007).

**F1a1** (Figure 5.34) is widespread in MSEA, ISEA, China, Japan and Taiwan. It dates to ~15 ka, with four substantial subclades: F1a1a, F1a1b, F1a1c, and F1a1d. Basal F1a1 lineages are seen in China (Zheng *et al.*, 2011), Japan (Tanaka *et al.*, 2004) and Peninsular Malaysia, and in HVS-I F1a1\* lineages are seen in Korea, Indonesia, Vietnam and Thailand. **F1a1b** dates to ~5 ka and is entirely a Japanese clade. **F1a1c** dates to ~11 ka with the basal lineages found in China (Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004). A separate subclade, defined by a transition at np 16224, is seen in one Northwest Peninsular Malay, dating to ~9 ka, and presumably dispersed from there into the Moken maritime minority off western Thailand (Pradutkanchana, Ishida and Kimura, 2010) as **F1a1c1**, dating to ~4 ka. **F1a1d** dates to ~5 ka and is reported in a single Northeast Peninsular Malay, a single Philippine individual (Tabbada *et al.*, 2010), and several aboriginal Taiwanese groups, especially the Yami, where it reaches 23% (Loo *et al.*, 2011), and where it further diversified as a subclade ~1.5 ka.

**F1a1a** is a common subclade dating to ~11 ka, and is widespread in MSEA and the Malay Peninsula (5.7% in Table 3.4; Figure 5.34). Basal lineages are found in single individuals from South China (Kong *et al.*, 2003b) and Sumatra (Gunnarsdóttir *et al.*, 2011b), with multiple individuals from Peninsular Malaysian *Orang Asli* and Malay (this analysis). The Semang Jahai (Jinam *et al.*, 2012) and Senoi Temiar share a subclade defined by a transition at np 6040 dating to ~4 ka; and another subclade within F1a1a1 with the same age. The derived cluster, **F1a1a1** dates overall to ~9 ka, and seen is only in Peninsular Malaysia and MSEA, where it is common; since it is not found in Indonesia, and given the prevalence of East Asian basal lineages in F1a1, an early Holocene migration from South China into

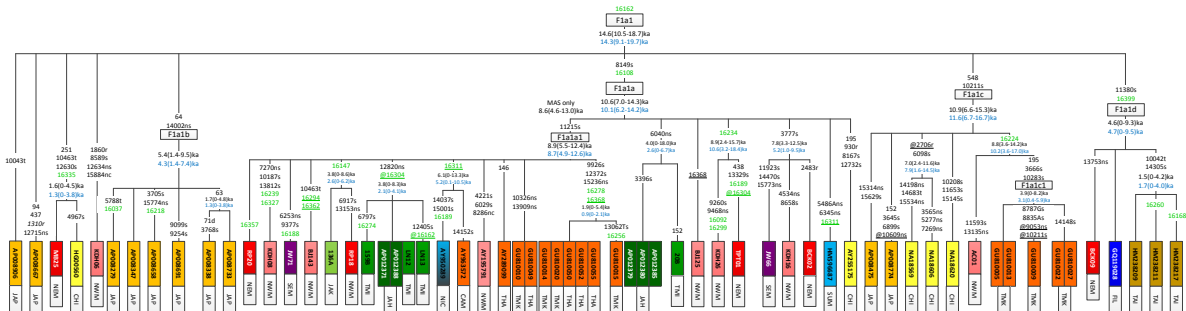
MSEA through Peninsular Malaysia might be indicated (matching the Malay/MSEA subclade of F1a1c), perhaps coinciding with the spread of the “coastal Neolithic” in MSEA (Higham, 2004; Bellwood, 2001).

A nested F1a1a1 subclade, defined by a transition at np 16147, dating to ~4 ka, and is seen in an Aboriginal Malay (Jahai) and a Northwest Peninsular Malay. As noted above, another F1a1a1 subclade, defined by a transition at np 12820 and a reversion at np 16304, is found in the Semang Jahai (Jinam *et al.*, 2012) and Senoi Temiar, dating to ~4 ka. The Nicobars (Thangaraj *et al.*, 2005) and Cambodia (Macaulay *et al.*, 2005) share another subclade defined by a transition at np 16311 dating to ~6 ka. Lastly, another subclade, dating to ~2 ka, consists of the Moken and Thai lineages (Pradutkanchana, Ishida and Kimura, 2010) only. Bellwood (1993, 1997) proposed a model for colonisation of SEA that simplified the number of migrations to 2, and explaining how such migrations may have occurred and their relation to language distribution. Southeast Asia’s negrito, including the Semang, would represent the relict descendants of SEA’s original “Australo-Melanesian” foragers. During the middle Holocene, both Austro-Asiatic and Austronesian languages arose in South China and were introduced to SEA with the Neolithic expansion of farmers (of Mongoloid physical appearance). Austro-Asiatic speakers took a mainland route southwards into MSEA, including Peninsular Malaysia and Nicobar Islands, whereas Austronesian speakers spread along the island arc from Taiwan to the Philippines, and then Indonesia and Malaysia. In the Malay Peninsula, interaction between immigrant farmers and resident foragers resulted in the mixed phenotype of certain groups, in particular the Senoi, as well as language shift by the Semang to Aslian (Hill *et al.*, 2006).

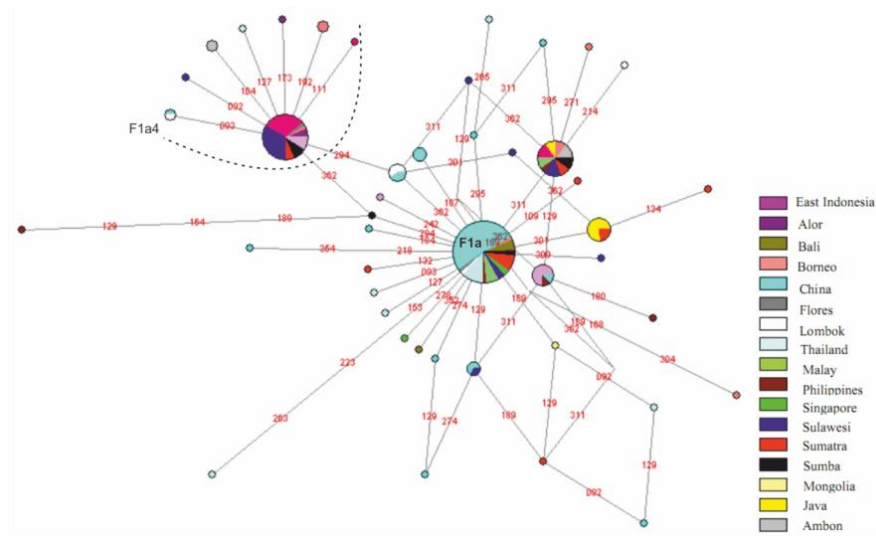
We can confirm the distribution found in the whole-mtDNA phylogeny by looking at the HVS-I networks. Figure 5.35 shows the HVS-I network of F1a\* (i.e., in HVS-I, F1a\* excluding F1a1). The root type of F1a is most common in Yunnan of South China. A derivative branch, F1a4, defined by a transition at np 16294, consists mostly of Island Southeast Asian types (Hill *et al.*, 2007).

Figure 5.36 shows the HVS-I root type of F1a1 remains most commonly found in South China and the Taiwanese Aborigines, followed by several individuals from Eastern Indonesia, Sumba, Borneo and Malay. Derivatives are found at low levels in Borneo, Sulawesi and Bali. One Malay individual falls on a branch with an additional transition at np 16335 with those from China and Singapore (represented in the whole-mtDNA tree, basal to

F1a1 in Figure 5.34). Another Malay falls in same branch with previously identified Malay individual with a transition at np 16224, which is identified as F1a1c now. F1a1\* types are rarely seen in ISEA and the result indeed complement the whole-mtDNA tree suggesting a possible Malay origin in Austronesian dispersal from South China and Taiwan.

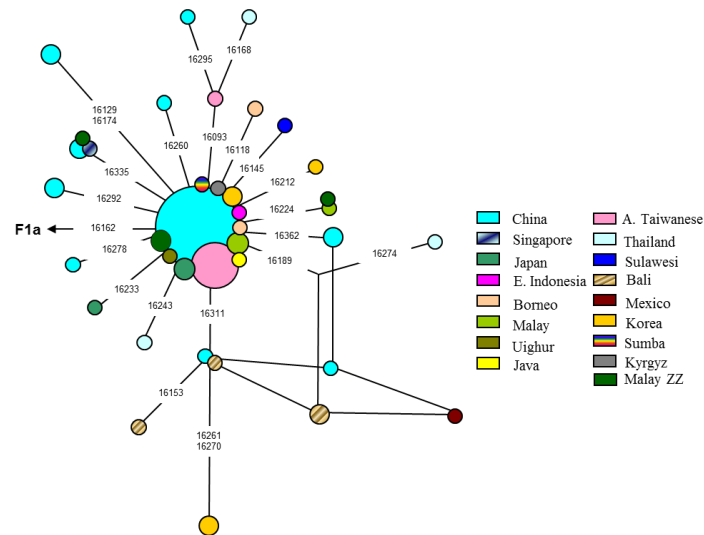


**Figure 5.34** The tree of haplogroup F1a1. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CAM – Cambodia, CHI – China, FIL – Philippines, JAH – Semang Jahai, JAK – Aboriginal Malay Jakun, JAP – Japan, NEM – Northeast Peninsular Malay, NIC – Nicobars, NWM – Northwest Peninsular Malay, SEM – Southeast Peninsular Malay, SUM – Sumatra, TAI – Taiwanese Aborigines, THA – Thailand, TMI – Senoi Temiar, TMK – Thailand Moken)

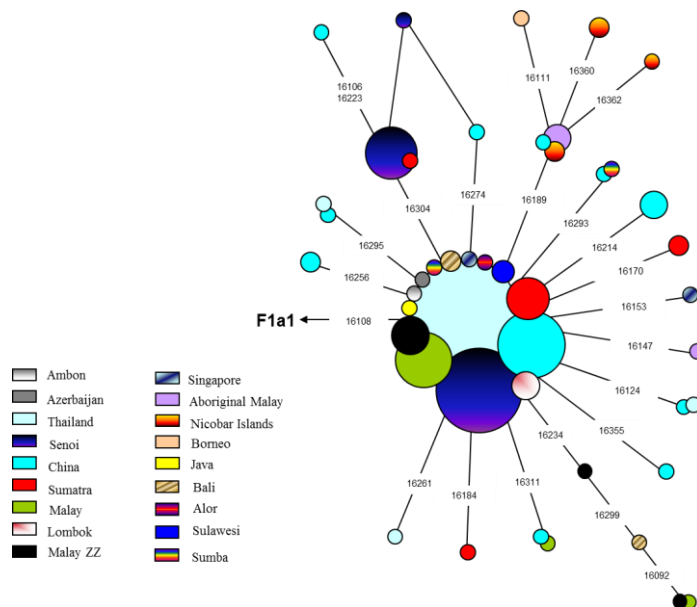


**Figure 5.35** HVS-I network of F1a1\*. F1a1 is further defined by transitions at nps 16129 and 16172 (van Oven and Kayser, 2009). Figure adapted from Hill (2005).





**Figure 5.36 HVS-I network of F1a1. Figure adapted from Hill (2005).**



**Figure 5.37 HVS-I network of F1a1a. Figure adapted from Hill (2005).**

F1a1a is widespread and relatively common in East and Southeast Asia (Figure 5.37), though it is most common in eastern MSEA (Mormina, 2007). In HVS-I, the root type of F1a1a is most commonly found in Thailand (~15% of the sample), Vietnam (~8%), coastal China (~4%), Peninsular Malaysia (the Senoi at ~40% and Malay ~8%), and ~3% in West Indonesia, and not found in Japan/Korea and Taiwanese Aborigines (Mormina, 2007). Its derivative types are detected in Sumatra, Bali, Sumba, Borneo and Peninsular Malaysia (the *Orang Asli* and Malay). One Peninsular Malay has an additional transition at np 16234, followed by a one-step derivative type from Bali with a transition at np 16299, and a further np 16092 in two Malay individuals (represented in the whole-mtDNA tree Figure 5.34).

Overall the pattern is similar to the distribution of basal lineages in the complete genome tree, although the HVS-I network suggests that China may be under-represented in the complete-mtDNA tree.

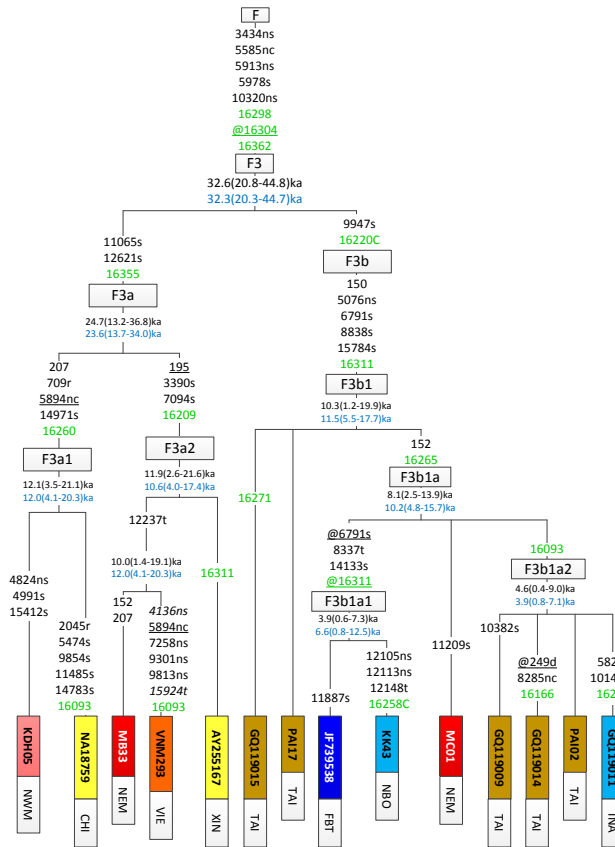
However, haplogroup F1a1a is most diverse in South China and MSEA (see Figure 67 in Mormina, 2007), along with its high levels of F1a\* and F1a1\*, suggesting an origin in South China or MSEA followed by migratory events into Peninsular Malaysia and, to a lesser extent, ISEA (see also Hill *et al.*, 2006). The whole-mtDNA F1a1a tree indicates an origin in South China followed by very rapid spread into MSEA in the early Holocene, coinciding with the “coastal Neolithic”. Interestingly, F1a1a appears to spread right down into Malaysia, rather than then hopping again in the mid to late Holocene into the Peninsula as proposed by Bellwood’s model for the spread of Austro-Asiatic speakers. It then seems to have entered the ancestors of the Senoi from within the Peninsula around 4 ka, which would correspond with the Bellwood model (1993, 1997) for the appearance of Austro-Asiatic in the timing that the Neolithic was brought into Peninsular Malaysia who intermarried with indigenous groups to create the ancestors of modern Senoi, but not the immediate source.

#### 5.6.4 Haplogroup F2

**F2** is an uncommon F branch, being predominantly found in China, with occasional instances in Japan and MSEA (Vietnam and Thailand), suggesting minor, recent movements. Detailed description is available in Appendix E.

#### 5.6.5 Haplogroup F3

**F3** dates to ~33 ka, and diverged into F3a and F3b (F3b1), estimated at ~25 ka and ~10 ka respectively (Figure 5.38). F3 has a wide distribution in China (Kong *et al.*, 2003b; Zheng *et al.*, 2011) and SEA. F3a sub-divides into F3a1 and F3a2, both of which are found in China and MSEA. **F3a1** dates to ~12 ka and is seen in one individual from China and a Northwest Peninsular Malay. The HVS-I database shows that it is present in South China and also, with very low diversity (and lacking the 16093 variant), in Vietnam and Thailand. **F3a2** dates to ~12 ka with a single basal lineage found in China and a subclade dating to ~10 ka in single individuals from Vietnam (Archaeogenetics Research Group, Huddersfield) and Northeast Peninsular Malay. The HVS-I database confirms its distribution in these three regions and also its extreme rarity.



**Figure 5.38** The tree of haplogroup F3. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, FBT – Philippines Batak, INA – Indonesia, NBO – North Borneo, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, TAI – Taiwan, VIE – Vietnam, XIN – Xinjiang, China)

Potential F3b\* lineages (lacking the 16311 variant) are found in several South Borneo individuals and a single South China individual in the HVS-I database. **F3b1** has basal lineages seen in Taiwan (Tabbada *et al.*, 2010), and indeed is common in the database exclusively amongst aboriginal Taiwanese. It diverged into **F3b1a** ~8 ka, with a basal branch seen in a Northeast Peninsular Malay. In the HVS-I data, F3b1a\* lineages are mainly found across Borneo, with single individuals in Taiwan aboriginals and Eastern Indonesia. F3b1a subsequently divided into F3b1a1 and F3b1a2, both of which are found in ISEA. **F3b1a1** dates to ~4 ka and is found in the Philippine Batak (Scholes *et al.*, 2011) (and the general Philippine population in the HVS-I data) and North Borneo. **F3b1a2** dates to ~5 ka, and is primarily found in Taiwanese aboriginals (as confirmed by the HVS-I database) and also a single individual from Indonesia (Tabbada *et al.*, 2010).

## 5.6.6 Haplogroup F4

**F4** dates to ~42 ka, and is divided into F4a1 and F4b (Figure 5.39). The age of **F4a1** is estimated at ~22 ka and further diverged into **F4a1a** ~4 ka in China (Ingman *et al.*, 2000; Hartmann *et al.*, 2009). A small clade nested below appears to have migrated by ~1 ka into Japan (Tanaka *et al.*, 2004). **F4b** dates to ~19 ka, with a deep-rooting lineage found in China (Zheng *et al.*, 2011) and a derived subclade **F4b1** ~10 ka, which appears to have spread into Northwest Peninsular Malay and (perhaps via there) India (Ingman and Gyllensten, 2003). The HVS-I database shows that F4b1 is found mainly in Taiwanese aboriginals, albeit with low diversity, and in several individuals in Sumatra, two of which share the 16170 variant with the Malay individual, pointing to a likely source for the Malay lineage seen here.

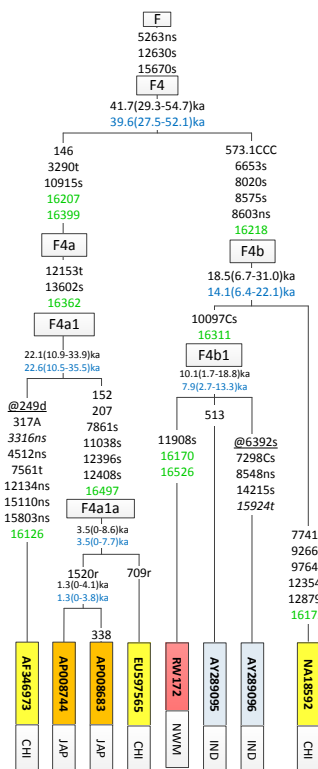


Figure 5.39 The tree of haplogroup F4. Time estimates shown for the clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, IND – India, JAP – Japan, NWM – Northwest Peninsular Malay)

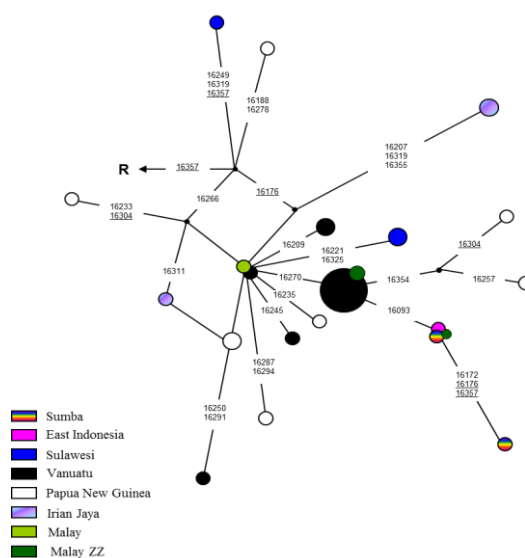
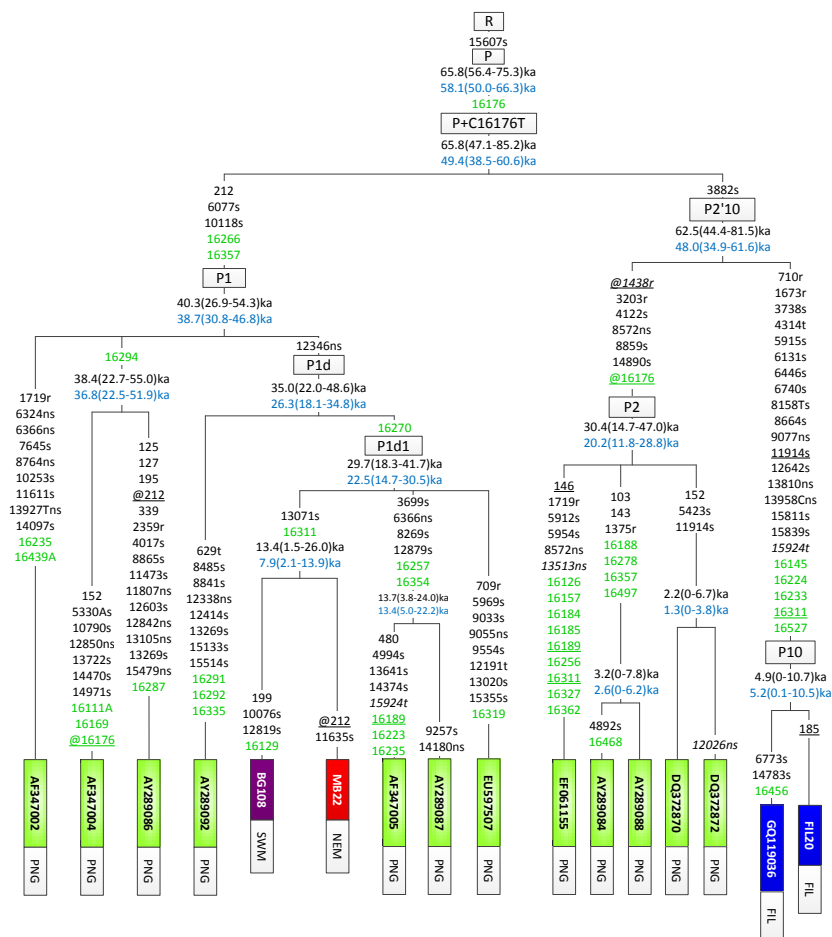
## 5.7 Haplogroup P

Haplogroup P dates to ~66 ka and is divided into P+C16176T (including P1 and P2'10), P3, P4, P5, P6, P7 and P9. Phylogeny P includes 33 complete sequences: 16 P+16176, five P3, seven P4, and five for P5, P6, P7 and P9. The entire haplogroup P has long terminal branches indicating genetic drift. Figure 5.40 shows haplogroup P and its major subclades

have deep ancestral roots and are widespread across Melanesia and Micronesia, Oceania. Several subclades that are found at low frequency elsewhere include P1d in the Peninsular Malay, and P9 and P10 in the Philippines. P3, P4a, P5, P6 and P7 are also found in the Australian Aborigines.

**Figure 5.40 Schematic diagram of haplogroup P and its major subclades distribution. (AUS – Australia, FIL – Philippines, INA – Indonesia, PEM – Peninsular Malaysia, OCE – Oceania)**

The HVS-I network in Figure 5.42 shows the branch defined by transitions at nps 16176 and 16266 (P1 in current nomenclature) is relatively common in Papua New Guinea and Vanuatu but not found in Micronesia. Elsewhere, haplogroup P is reported at low levels in East Indonesia, Sumba, Sulawesi and Peninsular Malaysia (Redd *et al.*, 1995; Hill, 2005; Zainuddin and Goodwin, 2004). As shown by the high-resolution whole-mtDNA tree and the HVS-I database, the deep ancestry is indeed found in Melanesia, it is almost certain that haplogroup P is an indigenous Melanesian haplogroup. Its presence in the Malay samples from Peninsular Malaysia and Indonesia indicates a certain degree of Melanesian contribution to the genetic make-up of ISEA.



**Figure 5.42 HVS-I network for P1. Figure adapted from Hill (2005).**

## 5.8 Haplogroup R6

**R6** is a rare haplogroup with deep root in India that dates to ~52 ka. Figure 5.43 shows the phylogeny is constructed with 8 complete sequences. **R6a** and **R6a1** date to ~52 ka and ~37 ka respectively, where the basal lineages are seen in Uttarpradesh, North India (Palanichamy *et al.*, 2004). **R6a1a** (dates to ~13 ka) and its subclade R6a1a1 (~3 ka) are in central and southeastern coast of India (Chaubey *et al.*, 2008; Sharma *et al.*, 2012).

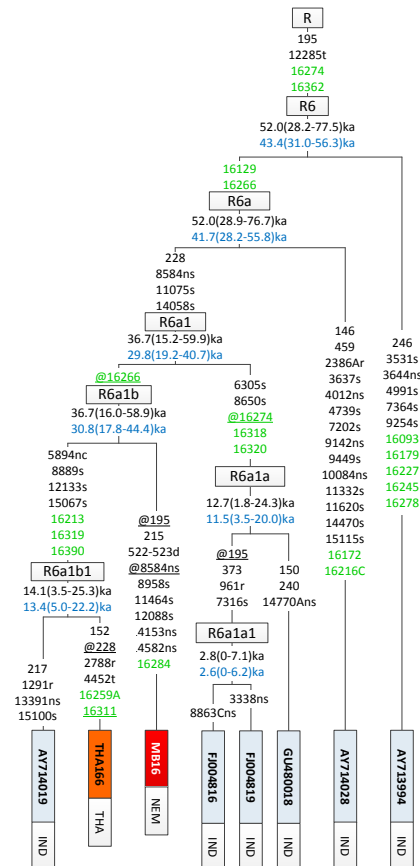


Figure 5.43 The tree of haplogroup R6. Time estimates shown for the clades are ML (in black) and averaged distance ( $p$ ; in blue) in ka. (IND – India, NEM – Northeast Peninsular Malay, THA – Thailand)

**R6a1b** dates to ~37 ka, where the basal lineage is present in the Northeast Peninsular Malay (this study). Subclade **R6a1b2**, dating to ~14 ka, is seen in South India (Palanichamy *et al.*, 2004) and Thailand (Archaeogenetics Research Group, Huddersfield). R6a1b shows early expansions from possibly South India into MSEA. The deep ancestry of R6a1b appears to be preserved in the Peninsular Malay population, and most of the lineages have undergone drift over time. However, it is known that the immigrants from India has arrived in the Malay Peninsula around the 17<sup>th</sup> century, this is likely to arrive recently.

## 5.9 Haplogroup R7

Haplogroup R7 dates to ~63 ka, and it is widely found in India, in particular the eastern coast (Figure 5.44). The basal lineage of R7 is seen in Andhra Pradesh of southeastern India (Fornarino *et al.*, 2009). The bigger clade nested within R7 dates to ~55 ka, and the basal lineage is interestingly found in Kota Kinabalu, North Borneo (Archaeogenetics Research Group, Huddersfield), suggesting an early arrival of modern human in SEA. **R7a'b**, dating to ~38 ka, is divided into R7a and R7b. **R7a** dates to ~18 ka, and R7a1 ~10 ka, where the basal lineages are mostly seen in India (Palanichamy *et al.*, 2004; Chaubey *et al.*, 2008; Sharma *et al.*, 2012), and one instance from Southwest Peninsular Malay (this study). **R7a1** has two subclades, R7a1a and R7a1b. **R7a1a** dates to ~4 ka, where it is found in Brazil, South America (Hartmann *et al.*, 2009) with a subclade nested within (~3 ka) in India (Chaubey *et al.*, 2008) and Pakistan (Fornarino *et al.*, 2009). Since it is only a single instance of Brazilian sample in a predominantly Indian haplogroup, this sample could be possibly migrated recently from India into South America. **R7a1b** dates to ~7 ka, and its subclades are confined to India (Chaubey *et al.*, 2008).

**R7b** dates to ~29 ka and it is divided into R7b1 (~18 ka) and R7b2 (~15 ka). The entire R7b is restricted to central and southeastern India (Palanichamy *et al.*, 2004; Chaubey *et al.*, 2008; Rani *et al.*, 2010). R7 clearly has a deep root in India, probably along the southeastern region with some offshoots arrived in North Borneo as early as ~55 ka and a Southwest Peninsular Malay lineage is found within R7a1, dating to ~10 ka. Considering the known history of Indian immigrants into Peninsular Malaysia during the 17<sup>th</sup> century, the Malay lineage is most likely to have arrived quite recently.





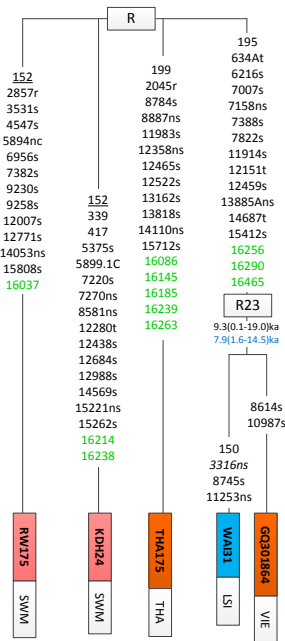


Figure 5.45 The tree of haplogroup R23 with three R\* lineages. Time estimates shown for the clades are ML (in black) and averaged distance (p; in blue) in ka. (LSI – Lesser Sunda Islands, SWM – Southwest Peninsular Malay, THA – Thailand, VIE – Vietnam)

## 5.11 Haplogroup R30

**R30** dates to ~66 ka and is divided into R30a and R30b. It is mainly found at low levels in South Asia. Detailed description is available in Appendix E.

## 5.12 Haplogroup U

Haplogroup U is uncommon in Asia but more diversified in Europe and the Near East, but the lineages seen in the Malays are largely those seen in Southwest and South Asia. Haplogroup U dates to ~61 ka and is divided into U1, U2'3'4'7'8'9, U5 and U6 (Figure 5.46). Haplogroup U is under-represented by 19 complete mtDNA sequences, mainly to show the phylogenetic relationships of the Peninsular Malay in the tree.

**U1a** dates to ~26 ka, which further diverged into **U1a1** ~20 ka, seen here in southern India (Ingman and Gyllensten, 2003) and Russia (Hartmann *et al.*, 2009), while U1a3 is represented by a single instance from the Northeast Peninsular Malay (this study).

**U2** dates to ~57 ka, and nested within is **U2b1** which is represented by an instance from the Northwest Peninsular Malay and **U2e1** in a Caucasian from North America (Mishmar *et al.*, 2003). **U4'9** dates to ~50 ka and found restricted to India (Fornarino 2009) and Pakistan

(Hartmann *et al.*, 2009). **U7a** dates to ~23 ka and U7a2 is found in Israel (Hartmann *et al.*, 2009) while U7a3 in Northwest Peninsular Malay (this study). **U5** dates to ~37 ka, and its subclades U5a (~27 ka) and U5b (~25 ka) are represented by instances from France, Italy, and Israel. **U6a1a** is seen in a single instance from Algeria (Hartmann *et al.*, 2009).

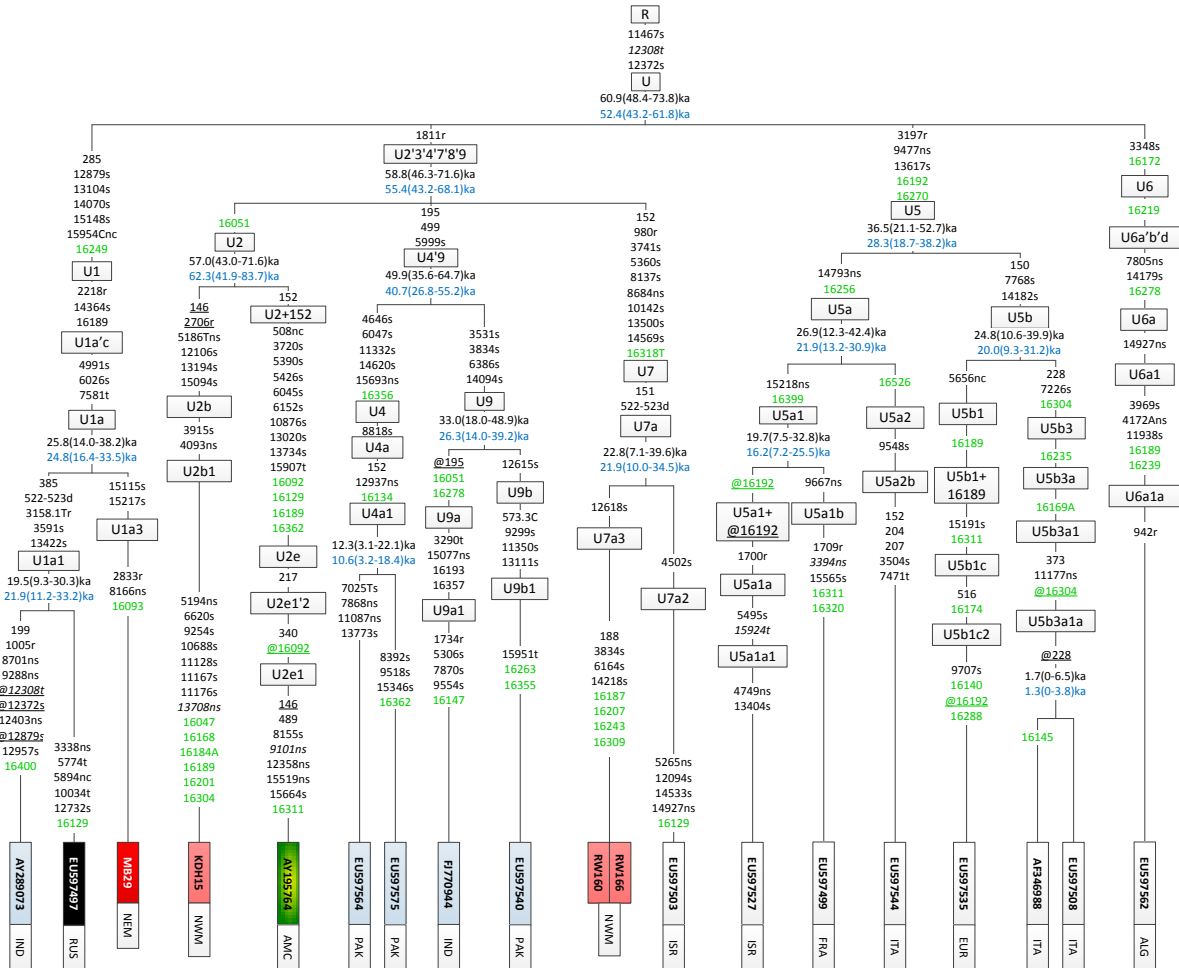


Figure 5.46 The tree of haplogroup U. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (ALG – Algeria, AMC – American, Caucasian, EUR – Europe, FRA – France, IND – India, ISR – Israel, ITA – Italia, NEM – Northeast Peninsular Malay, NWM – Northwest Peninsular Malay, PAK – Pakistan, RUS – Russia)

## 6 Bayesian Skyline Plot (BSP)

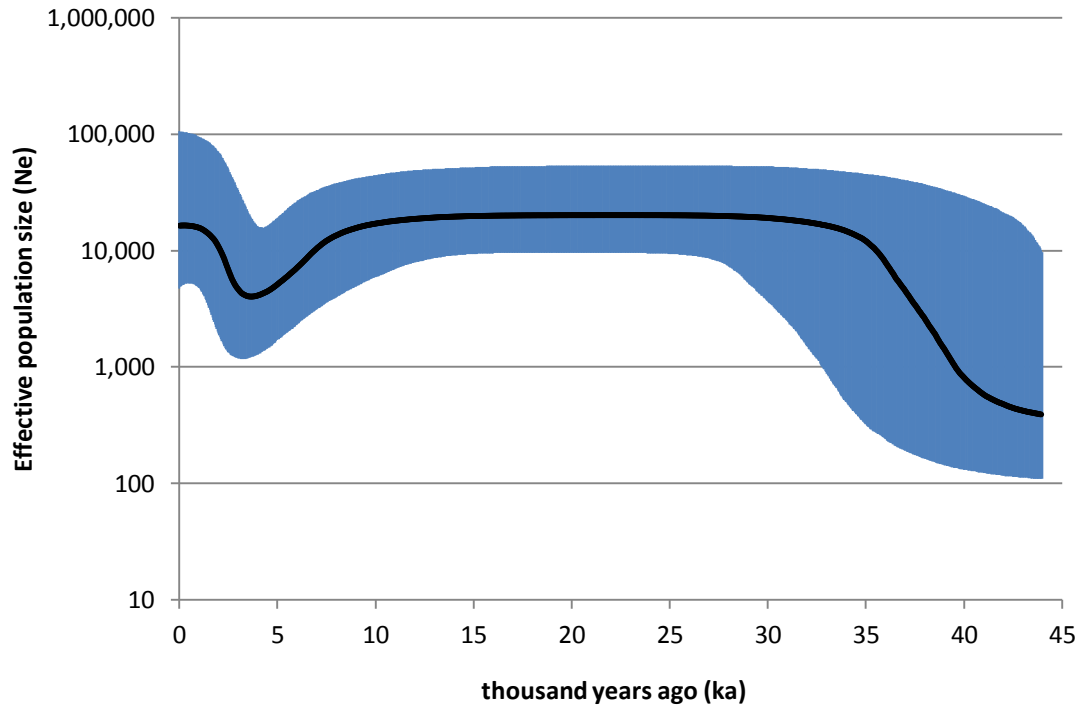
Periods of major expansion (population growth) can be reconstructed using Bayesian Skyline Plots (BSPs), using the Bayesian evolutionary analysis by sampling trees (BEAST) software package, when the data are sufficiently informative about the population (Drummond *et al.*, 2007). See more in Section 2.9.3.

Recent BSP studies in SEA include the Philippine populations ( $n = 92$ ) (Gunnarsdóttir *et al.*, 2011a), Malaysian indigenous people ( $n = 86$ ) (Jinam *et al.*, 2012), and Indonesia ( $n = 2104$ , but unusually using control-region, rather than whole-mtDNA, sequences) (Guillot *et al.*, 2013), where a characteristic general pattern of growth during the Pleistocene and recent decline was detected. In the Philippines, the BSPs using whole-mtDNA genomes for the Mamanwas (negrito), Manobos and Surigaonons indicate population growth from ~50 ka until ~30-35 ka, followed by population stasis until ~6-8 ka, at which point population size decreases. Additionally, the Surigaonons differ from the other groups in showing another signal of population growth beginning ~2-3 ka (Gunnarsdóttir *et al.*, 2011a).

Jinam *et al.* (2012) generated BSPs using coding-region mtDNA sequences from the Malaysian indigenous populations (Semang Jahai, Aboriginal Malays Temuan and Seletar, and Austronesian-speaking Bidayuh in Sarawak). They observed an increase in population size ~60-40 ka and stasis from 30-10 ka, followed by a decline which lasted until several hundred years before present. Their BSPs also showed a slight increase of population size in all four groups after ~1 ka; but a possible cause was not offered in the study (Jinam *et al.*, 2012).

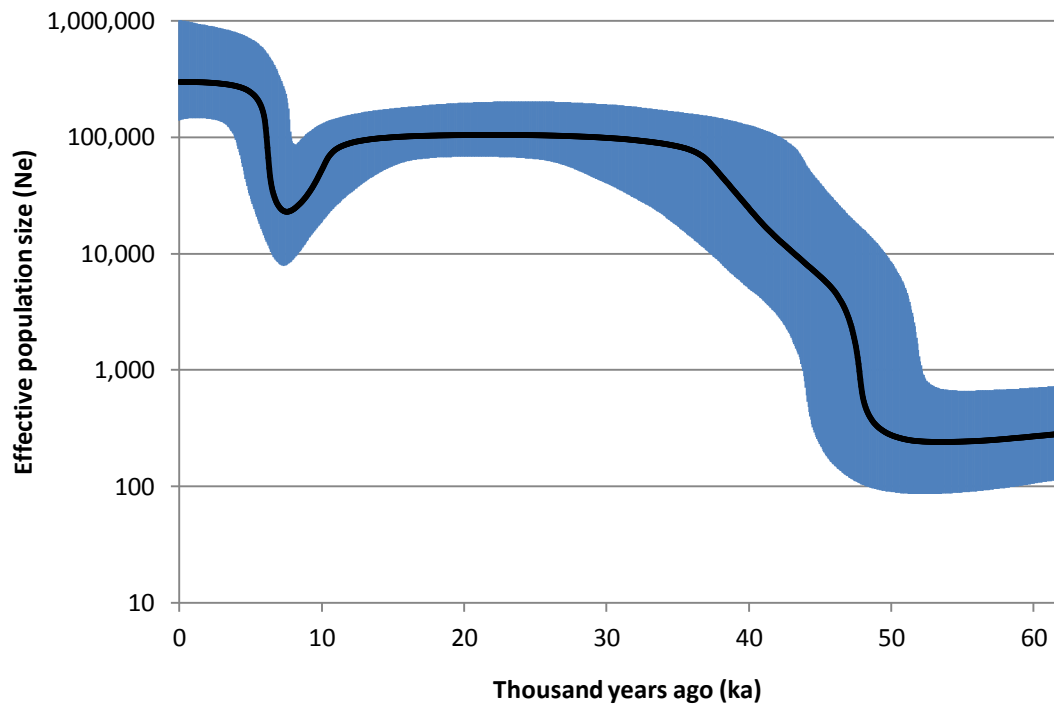
Guillot *et al.* (2013) used low resolution HVS mtDNA sequences but large sample size ( $n = 2104$ ) from four islands of the Indonesian archipelago (Bali, Flores, Sumba and Timor) and found little evidence for large fluctuations in effective population size. Their studies found a slow population growth during the late Pleistocene that peaked 15-20 ka, with subsequent slow decline into the Holocene. They suggested that this pattern may reflect population declines caused by the flooding of lowland hunter/gatherer habitat during sea-level rises following the LGM. The weaker signal may reflect the lower resolution of the data, however the same inference could apply to the early Holocene decline in the Philippine populations noted by Gunnarsdóttir *et al.* (2011a).

Figures 6.1 and 6.2 show my BSPs in Peninsular Malaysia. I first performed the BSP analysis on the *Orang Asli* populations, including all six Semang sub-groups (Kensiu, Kintak, Batek, Jahai, Mendriq and Lanoh), two Senoi (Semai and Temiar) and three Aboriginal Malay (Jakun, Semelai and Temuan). The second BSP includes all 186 modern Malay complete sequences in this study.



**Figure 6.1 Bayesian skyline plot (BSP) indicating hypothetical effective population size over time of *Orang Asli* populations. The posterior effective population size through time is represented by the black line. The blue region represents the 95% confidence region. Effective population size is plotted on a log scale.**

There is a significant expansion ~49-fold in the *Orang Asli* plot between ~44 ka to ~30 ka (Figure 6.1), likely to coincide with the emergence of subclades M22a, M13b1 and N22. The effective population size peaks at the LGM, ~22 ka, with a plateau between ~35 ka and ~15 ka. It then shows an early Holocene crash falling to about a quarter of the peak size by the mid-Holocene, indicating a major bottleneck, and starts to rise again after ~4 ka with the final effective population size restored to almost to the same as during the Late Pleistocene. The recent rise fits well with the emergence of haplogroups M21a1b in the Semang and Aboriginal Malay; N22a in the Aboriginal Malay Temuan; N21a1a in Temuan and Semelai; N9a6a in the Semang Jahai, Kensiu, the Aboriginal Malay Temuan and Seletar; R21a1 in Semang and Senoi Temiar; and R9b1a1a3 in Kintak, Jahai and Aboriginal Malay.



**Figure 6.2 Bayesian skyline plot (BSP) indicating hypothetical effective population size over time of Peninsular Malay populations. The posterior effective population size through time is represented by the black line. The blue region represents the 95% confidence region. Effective population size is plotted on a log scale.**

Both plots show, what is presumably the same expansion from the same ancient Sunda source in Peninsular Malaysia, but the signal is better preserved in the modern Malay data due to the smaller impact of recent genetic drift, which has been extremely pronounced in some *Orang Asli* groups with very small population sizes, especially the Semang and Senoi (Carey, 1976). Figure 6.2 shows the initial Palaeolithic expansion of the Peninsular Malay, occurring from 50 ka to 36 ka with an estimated 277-fold increase – a signal of the first colonisation of Southeast Asia during the African exodus (based on the phylogenetically more complete Malay data; for the *Orang Asli*, the increase is ~49-fold between 44 ka and 30 ka). This is likely shown by the emergence of ancient haplogroups for examples M4'67, M13b, M21c'd, M50, M17c, M12, M26a, M1'20'51, M71, R22, R7, F, B4c and B5a'b.

We need to remember that the *Orang Asli* sample is made up of a number of very different sub-groups that are conflated into a single BSP. The effects of very recent crashes are probably minimized because different *Orang Asli* groups preserve different fragments of the overall diversity. Similar effects are possibly detected by Jinam *et al.* (2012) but the small dips appeared to be much more recent, ~1 ka (See Figure 6 in Jinam *et al.*, 2012). On the other hand, the modern Malay preserve much more, including lineages from ISEA not seen in

the *Orang Asli*, as well as indigenous Malay Peninsula lineages that have been lost by drift from the *Orang Asli*.

Both also show quite ancient crashes in the Holocene, which was not seen previously in lower-resolution studies in human populations of SEA (Gunnarsdóttir *et al.*, 2011a; Jinam *et al.*, 2012; Guillot *et al.*, 2013). The Malay crash during the early Holocene ~11 ka, recovering in the mid-Holocene ~5 ka; the *Orang Asli* probably show the same signal, but with less clarity, due to more recent drift effects and possibly also substructure. The crash was predicted from the long branches evident in haplogroup E in ISEA (Soares *et al.*, 2008) and the re-expansion probably represents the mid-Holocene starbursts that we see *e.g.* in haplogroups E and B4a1a. The crash (also unique to the Sunda populations) coincides with the sea-level rises over the huge Sunda shelf – the population peak is at the LGM, 22 ka, and there is a dramatic rise of ~13-fold from 7.5 ka, with the final effective population size about ~3-fold the size it was during the Late Pleistocene. This fits well the model of Oppenheimer (1998) and Soares *et al.* (2008) that proposes an initial catastrophic effect on the people of the region between the LGM and the final sea-level rise (~7.5 ka) and a major recovery as some populations re-adapt to coastal living and expand along the extended coastlines now available.

## 7 Discussion and Conclusions

The ancient continent of Sundaland (today's Mainland Southeast Asia, Sumatra, Borneo and Java) is thought to have been one of the key areas of primary settlement of the anatomically modern humans who dispersed along the coastal route of Indian Ocean into Asia and Australia. Previous mtDNA studies on relict populations from the Malay Peninsula identified novel basal M, N and R lineages in the *Orang Asli* groups in Peninsular Malaysia, suggesting these three founders moved along the south coast of Asia ~50–60 ka, reaching Southeast Asia and the Sahul continent (Australia and New Guinea) by ~50 ka. It seems likely that these deep rooted haplogroups found in Peninsular Malaysia can be traced back to the original inhabitants of Southeast Asia, who first colonised the Sunda area ~50–60 ka.

Semang and Senoi differ greatly in the extent and composition of Holocene genetic introgression from outside, although they share several indigenous Pleistocene mtDNA lineages (M21a, M13b, and R21). The ancestors of the Semang negrito are generally supposed to be the least changed in all respects, physical and cultural, while the ancestors of the Senoi adopted hill rice farming. The Aboriginal Malays have distinct and indigenous Pleistocene founding mtDNA lineages different from Semang/Senoi (such as M7c3c, M22, N21, and N22) and tend to horticulture rather than rice farming (Oppenheimer, 2011).

The ancestry of the lineages is shown by the assignment of *Orang Asli* lineages (using control-region data) to putative proximal source regions on the basis of standard phylogeographic principles, so that a majority of basal (or deep) lineages within a cluster in one region was taken to indicate that that region was the likely source for the cluster. For example, the deeper branches in M17 (indeed almost all the lineages) are largely restricted to MSEA, the Malay Peninsular and western Indonesia, suggesting a Sunda source.

I divided the putative sources into three regions: East Asia, ISEA and MSEA/Sunda (Table 7.1). Table 7.1 shows that ~89% of the *Orang Asli* lineages are most likely indigenous to MSEA/Sunda shelf, dating between the late Pleistocene to the early Holocene (~50–8.5 ka). ~7% of the total lineages appear to illustrate gene flow from ISEA around the mid-Holocene, between ~8.5 ka to ~2.5 ka. Haplogroups B4a1c and M7c1a, representing ~0.7% of the lineages, can be traced back to East Asia. Lastly, ~4% consist of M\* and B\* that are not able to be assigned to a source because of the low resolution of the HVS-I data.



**Table 7.1 Assignment of *Orang Asli* lineages to putative proximal source regions. Figures taken from Table 3.3. (EA – East Asia, ISEA – Island Southeast Asia, MSEA/SUN – Mainland Southeast Asia/Sunda)**

Haplogroup	Putative proximal source	Semang						Senoi						Aboriginal Malay/Proto-Malay						Total	%
		Batek	Jahai	Lanoh	Kensiu	Kintak	Mendriq	CheWong	Jah Hut	Mah Meri	Semai	Senok Beri	Temiar	Jakun	Kanak	Kuala	Seletar	Semelai	Temuan		
B4a1c	EA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0.5
M7c1a	EA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0.2
<b>Total East Asia</b>																					<b>0.7</b>
B5b1c	ISEA	13	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	15	3.4
E	ISEA	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	3	0	0	6	1.4
M7c3c	ISEA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	9	2.1
<b>Total ISEA</b>																					<b>6.9</b>
B4c2	MSEA/SUN	0	0	0	0	2	0	0	0	0	1	0	0	1	0	0	0	0	0	4	0.9
B5a	MSEA/SUN	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	3	0.7
F1a1a	MSEA/SUN	0	5	6	5	5	0	0	0	5	6	5	37	1	0	0	0	4	1	80	18.3
M13b	MSEA/SUN	0	2	0	0	1	2	0	0	0	0	0	1	0	0	0	0	4	2	12	2.8
M17a1a	MSEA/SUN	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0.5
M21a	MSEA/SUN	19	9	5	11	15	31	5	5	0	0	0	5	0	0	0	0	2	2	109	25.0
M21c	MSEA/SUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0.5
M22	MSEA/SUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	6	1.4
N9a6	MSEA/SUN	0	9	0	1	1	0	0	0	0	0	0	4	0	0	0	2	1	4	22	5.0
N21	MSEA/SUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	7	27	6.2
N22a	MSEA/SUN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	1.1
R9b	MSEA/SUN	0	2	0	0	2	0	0	0	0	0	0	0	1	0	0	0	17	7	29	6.7
R21	MSEA/SUN	1	33	4	18	2	2	0	0	0	0	0	22	1	0	0	0	2	0	85	19.5
<b>Total MSEA/SUN</b>																					<b>88.5</b>
B*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0.7
M*	Unknown	0	0	0	0	0	0	0	0	0	0	0	2	1	4	5	0	1	1	14	3.2
<b>Total unknown</b>																					<b>3.9</b>
<b>Total</b>		<b>34</b>	<b>60</b>	<b>15</b>	<b>37</b>	<b>28</b>	<b>37</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>7</b>	<b>5</b>	<b>72</b>	<b>7</b>	<b>5</b>	<b>5</b>	<b>5</b>	<b>66</b>	<b>38</b>	<b>436</b>	<b>100.0</b>

## 7.1 Semang and Senoi

In order to establish whether the Semang are the descendants of the earliest settlers, it is helpful to look at the extent of the indigenous Semang lineages not found elsewhere, as opposed to lineages shared with other populations. A very small component of these lineages at ~0.5% is found only in the Semang, amongst the *Orang Asli* groups – for instance haplogroup M17a1a. M17a1a dates to the late glacial period ~15.5 ka and is found in Vietnam, with a branch dating to ~2.4 ka in the Semang. The rest of the Semang lineages, including their most frequent haplogroups M21a and R21 (below), are shared variously with Senoi and Aboriginal Malays (as well as some modern Malay) indicating some degree of genetic admixture amongst the *Orang Asli* groups.

Finding Semang, Senoi, Aboriginal Malays and Malay in M13b1 within haplogroup M13b and alongside other MSEA and ISEA types, which are found on all primary M13'46'61 branches, might indicate an overall MSEA/Sunda origin for the whole M13'46'61 clade dating to the Pleistocene ~58 ka. Two of the recent offshoots indicated by subclades M13b2 and M61a have expanded into India and Tibet respectively at ~6.5 ka and 4.1 ka (more discussion below). The whole-mtDNA tree of M13'46'61 (~46 ka) indicates an ancient period on the Sunda shelf before South Asian populations began to differentiate significantly.

M21 and R21 were the first predominant, basal *Orang Asli* haplogroups identified by Macaulay *et al.* (2005). However, it has become clear that, unlike originally supposed, these clades are not unique to the *Orang Asli*, as they are found at lower rates elsewhere in Southeast Asia. Nevertheless, M21 was thought by Hill *et al.* (2006) to represent a component of deep Upper Pleistocene ancestry within the Malay Peninsula, and this study confirms that haplogroup M21 has a deep Sunda origin (~60 ka), as it is mainly found in the Malay Peninsula/South Thailand, with only a few sporadic cases further afield.

My whole-mtDNA study also confirms HVS-I evidence suggesting that the Semang and Senoi remain confined to and predominant (17/24) within the largest clade (M21a1b), which dates only to ~6 ka, in Peninsular Malaysia, with five instances of the same haplotype in Aboriginal Malay within the largest subclade of M21a1b (Figure 3.8). The remainder of the M21a diversity is rather preserved within the Malay (and one Aboriginal Malay) and adjacent populations, with the earliest branch seen in a single Philippine individual. Only a small

subclade of M21a1b, and dating ~3 ka, is restricted solely to Semang (including Batek, Lanoh, and Jahai). Meanwhile, the second, larger subclade of M21a1b, dates to ~3.5 ka and, with two non-OA MSEA exceptions (a modern Malay and a Thai), is found only in Semang (Kensiu, Kintak, Mendriq and Jahai) and five Aboriginal Malay Temuan. Complete mtDNA sequence M21c was only found in two Semelai and is not readily recognisable as a clade by using HVS data only and so it was not possible to ascertain the distribution of M21c further with HVS data.

The most likely explanation for this pattern in the ‘young’ peninsular M21a1b cluster within M21a1 is local genetic drift among ancestors of the Semang, shortening the coalescence time, and recent gene flow from Semang and/or Senoi into the Aboriginal Malay Temuan, as well as into the modern Malay and Thai, rather than a local Holocene founding event. However, M21a1 (~14 ka), along with M21c and M21d, have a much wider distribution centred on the Sunda region, seen in the modern Malay, MSEA and South Borneo, extending in several sampled individuals to the Philippines and Sulawesi. The ancient ancestry of M21a (and M21 as a whole) therefore appears to be more fully captured by in other “non-relict” populations around the Sunda shelf, indicating a Pleistocene Sunda origin rather than that M21 first spread out from *Orang Asli* populations. The significance of the predominance of M21 among the negrito Semang, when compared with other *Orang Asli* groups and Sunda as a whole, may simply reflect isolation and less admixture with later influxes.

R21 is present (to very different extents) in all six Semang groups and predominant in two out of six, but it is present (and common) in only one out of six Senoi groups and present at only low rates in two out of six Aboriginal Malay groups. R21 was previously undated, but shown to diverge basally (or almost so) from the ancestor of haplogroup M to ~60 ka using coding-region data by Macaulay *et al.* (2005) and thought to be a possible sister clade of haplogroup R9 by sharing a transition at np 16304 (Hill *et al.*, 2006). However, the new whole-mtDNA phylogenies, including this study, show that R9 and R21 belong to different haplogroups. This study dates R21 to ~12 ka with R21a (at ~ 6 ka) remaining highly localised within the northern Semang, especially the Jahai and in Senoi Temiar groups (Figure 5.26). However, one Malay complete sequence lineage has been found in this study, basal to R21, which has preserved a deeper ancestry no longer seen in the extant *Orang Asli* population, resulting in the deeper age estimate for R21 as a whole. Even so, whilst Hill *et al.* (2006)

identified R21 as a component of deep Upper Pleistocene ancestry within the Malay Peninsula, my study shows that sampled R21 lineages coalesce only at the end of the Pleistocene (~12 ka) presumably as a result of genetic drift and population subdivision of the *Orang Asli*, as with M21a. However, R21 appears to share a very deep ancestry with the Aboriginal Australian haplogroup R12, linking the Sunda and Sahul shelves at high time depth (an especially imprecise estimate, but >44 ka).

The cultural counterparts of M21 and R21 are suggested by Bulbeck (2011) to be the core- and cobble-based stone tool assemblages that led to ‘bifacial Hoabinhian’ assemblages in Peninsular Malaysia by 15 ka. M22 appears to have a Pleistocene source in MSEA/Malay Peninsula, dating back to ~50 ka, and the Aboriginal Malay Temuan and Peninsular Malay are found within subclade M22a2 (Figure 3.11). Bulbeck (2011) suggested that M22a2 may spread down the Peninsula under the same cultural category (the core- and cobble-based stone tool assemblages) with M21 and R21 even though it is found only in Temuan.

The distribution amongst the Temiar is very intriguing. It is worth noting that the Jahai share settlements with the Central-Aslian-speaking Temiar and they are in frequent contact with the modern Malay (Burenhult, 2001), which may be reflected by the evidence in this study within subclade R21a.

N9a has a much more widespread distribution. It is common across East Asia and Taiwan; however, subclade N9a6 is entirely restricted to MSEA, Peninsular Malaysia and ISEA, and entirely absent in Taiwan and the Philippines. N9a6 appears to have an origin in MSEA, with dispersals southwards through the Sunda shelf during the Late Glacial period ~16 ka, with deep branches in Malay, Aboriginal Malay and Sumatra and a derived subclade, within N9a6a, comprising Semang and Senoi dating to less than 5 ka. This broadly substantiates a previous study based on the HVS-I analysis suggesting that N9a6 had a Holocene intrusion from Indo-China (Hill *et al.*, 2007), although the details remain elusive.

The age of the Semang N9a6a subclade dates to ~5 ka, might suggest a correlation with the early Neolithic in Peninsular Malaysia, although the correlation suggested by Bulbeck (2011) was based on earlier estimates (based on Hill *et al.*, 2007) that are revised substantially upwards here. Bulbeck suggested a source for the Peninsular Malaysia’s early Neolithic in North Vietnam’s Bacsonian (11,000-7,000 BP) and/or Dabutian (6,500-4,500 BP) sites, pointing to a linked introduction of the N9a6 haplogroup and the Neolithic from

North Vietnam to Peninsular Malaysia (Bulbeck, 2011). The deeper time depth estimated here suggests rather an earlier, Late Glacial dispersal south.

B5a as a whole represents ~10% of the sample in MSEA and among the modern Malay and the highest frequency in China is in Southwest and coastal China at ~7% (Mormina, 2007). The diversity of this clade is also at its highest in MSEA, suggesting an origin in MSEA during the LGM, ~22 ka. B5a1 shows a post-glacial Sunda distribution with recent offshoots. B5a1a and its subclades illustrate a starburst expansion ~8 ka (Fig 5.18) clearly centred on MSEA and Peninsular Malaysia, with one branch extending into the Nicobar Islands at ~4.3 ka. B5a1a may therefore correspond to a hunter-gatherer dispersal that took place in the early Holocene, correlating with Bulbeck's (2008) hypothesis of the expansions of the coastal Neolithic Da But culture from around 6,000 BP (see also Higham, 2004).

A similar early Holocene starburst with a northern source is clearly seen in haplogroup B5b1, which has a likely origin in South China ~28 ka, while its derived subclade B5b1c dates to the early Holocene ~11 ka, and has a very distinctive distribution in the Sunda region, found only in the Semang Batek, Peninsular Malay and the Philippines (Figure 5.20). Haplogroup B5b1c in Semang and Malay may again represent descendants of the coastal Neolithic hunter-gatherers surviving on the peninsular remnant of the Sunda shelf after the second rapid sea level rise (~11.5 ka), in this case linking two distinct negrito populations – the Austro-Asiatic-speaking Semang Batek and the Austronesian-speaking Philippines Mamanwa, on the other side of the Huxley Line (see Figure 23 in Oppenheimer, 1999), within the same clade; a link not found in previous studies (e.g. multidimensional scaling analysis by Heyer *et al.*, 2013). Unfortunately, a more precise phylogeographic link between the negrito populations cannot be established because of the gaps present in the Philippine sequences (Gunnarsdóttir *et al.*, 2011a).

Haplogroup B4c2 is another lineage cluster that encompasses all three *Orang Asli* groups (Figure 5.15) as well as modern Malay, Sumatra and MSEA, with the centre of gravity of B4c pointing to deeper ancestry to the north. Two Semang Kintak, a Senoi Semai, an Aboriginal Malay Seletar and four southwest Peninsular Malay all occur in different positions within this haplogroup, which again appears to have an early Holocene ancestry within MSEA/Sunda dating to ~8.5 ka. The Senoi Semai and Aboriginal Malay Seletar share a younger clade of B4c2b (dating to ~4 ka) with three modern Malay and two Vietnamese. This possibly again indicates an eastern coastal Neolithic expansion from MSEA down the

Peninsula. Similarly, it has been suggested that F1a1a in the Senoi suggests the same immigration route correlating with coastal Neolithic expansion across the Gulf of Siam into Peninsular Malaysia via the Isthmus of Kra, perhaps with several entries (Oppenheimer, 2011, discussed more below).

The Semang and Senoi therefore share a common, predominantly local population ancestry but the ancestors of the Senoi have evidently undergone quantitatively more genetic and cultural admixture from outside than the Semang. This is primarily from Indo-China (Fix, 2011; Burenhult *et al.*, 2011; Oppenheimer, 2011), except for the intrusive B5b in the Batek and Mendriq Semang (Table 7.1). The Semang overall show less genetic intrusion than the Senoi, and paralleled by less change of lifestyle and physical morphology. The Senoi, although of a similar physical nature to Semang, tend to show plesiomorphic physical traits and have much higher rates of genetic intrusion from Indo-China, as illustrated by haplogroups N9a6, F1a1a, B5a1a and B4c2b (also see Oppenheimer, 2011). Like the Semang, the Senoi populations do not seem to show any recent mtDNA genetic source from India nor from any other region outside MSEA (Table 7.1), either of which might be expected according to the traditional and oversimplified layer-cake model.

There are lineages apparently indigenous to the southern Peninsula among the Senoi, like M13b1, consistent with the local differentiation model (Rambo, 1988; Benjamin, 1985, 1986). Other lineages among the Senoi are intrusive from Indo-China (i.e. northern MSEA). In my new whole-mtDNA phylogeny (Figure 5.34), haplogroup F1a1a in most of the Senoi groups (as well as, to a lesser extent, in the Semang, Aboriginal Malay and Malay) appears to have a source in northern MSEA, possibly further up the Mekong. As previously suggested (Hill *et al.*, 2006), this lineage may have accompanied the movement of the Aslian-languages into the Malaysian Peninsula, possibly along the Gulf of Siam via the Isthmus of Kra (Oppenheimer, 2011). Haplogroup N9a6 appears consistent with a northern source and Late Glacial dispersal, while several haplogroups point to a later, coastal Neolithic expansion from northern MSEA, possibly coinciding with the spread of the Da But culture as suggested by Bulbeck (2008).

### **7.1.1 The “Negrito Hypothesis”**

The negrito hypothesis anthropologically categorised various contemporary groups of hunter-gatherers in Southeast Asia, in particular the Andaman Islands, the Malay Peninsula,

and the Philippines, who share the phenotypes of dark skin, short stature, and tight curly hair. The shared phenotypes could be due to a common descent from a region-wide, pre-Neolithic substrate of humanity, as the hypothesis suggests, or alternatively, convergent evolution. Since the hypothesis remains unproven, it has been suggested that the designation be presented lowercase, as “negrito” avoid creating a possibly spurious unity amongst potentially disparate groups of people (Endicott, 2013).

Under this hypothesis, language shifts must have occurred at some time, except possibly in the case of the Andamanese. All Philippine negritos speak Austronesian languages similar to those of other Philippine populations, and all Malaysian negritos (the Semang) speak languages in the nuclear Mon-Khmer branch of Austro-Asiatic. The Andamanese remain distinct, showing possible limited affinity with a few small isolates in New Guinea and eastern Indonesia, and no widely accepted interpretation of the relationship of the Andamanese languages to the extant linguistic families of the South Asian region (Blevins, 2007). Blust (2013) suggested, in favour of the Negrito hypothesis, a common cultural and linguistic past for the Malaysian and Filipino negrito populations at a time which probably preceded the end of the Pleistocene, with the Andamanese possibly separating earlier.

Although a number of linguists have been in favour, osteological and population genetic studies have provided little evidence for the Negrito hypothesis. For example, Stock (2013) found no differences between the stature of Andaman Islanders and Filipino Aeta foragers in relation to phenotypic variation among hunter-gatherer groups more globally. Bulbeck (2013) found Andamanese and Semang (and Senoi) people to be osteologically more similar to each other, while Philippine negritos were dissimilar to both.

Both Chaubey and Endicott (2013) and Jinam *et al.* (2013) have studied the negrito populations using genome-wide autosomal SNP data and found relatively recent admixture from adjacent regional populations. They found some possible ancestral links between some of the groups, but no evidence of a single ancestral population for all of the different groups traditionally defined as ‘negritos’ in Southeast Asia. Various phylogeographic studies of the negrito populations using mtDNA and Y chromosome have found unique haplogroups in each negrito population, but none in common between them. For instance, Y-chromosome haplogroups C-RPS4Y and K-M9 (Delfin *et al.*, 2011) and mtDNA lineages B4b1 and P9 (as well as P10) are found in the Philippine negritos (Heyer *et al.*, 2013), M31 and M32a in the

Andamanese (Thangaraj *et al.*, 2005, 2006), and M21, M22 and R21 in the Semang (Hill *et al.*, 2006, 2007).

Moreover, McAllister *et al.* (2013) analysed the mtDNA haplogroups by SNP hierarchical typing of short-statured Australian Aboriginal groups in Far North Queensland (FNQ) and Tasmania and found that they carry lineages found in other Aboriginal Australian groups, and not those found in Southeast Asian negritos. Their result coincides with two Y chromosome haplogroups, C-RPS4Y and K-M9 in the Filipino negritos that are also shared with indigenous Australians (Delfin *et al.*, 2011).

This study also confirmed previous findings in Hill *et al.* (2006) that two haplogroups predominantly found in the Semang, M21a and R21, are not found in the Philippines, nor among Andamanese negritos. Intriguingly, a Batak negrito from the Philippines is seen to carry a lineage within the M21c1 subclade, which dates to ~31 ka, and which is shared with another Filipino, a Lesser Sundanese, an Aboriginal Malay Semelai and modern Malay in Peninsular Malaysia, but *not* with any Semang (Figure 3.9). Haplogroup B5b1c dates to soon after the second flood ~11 ka, incidentally linking a cluster of Austro-Asiatic-speaking Semang Batek and the Austronesian-speaking Philippine Mamanwa within the same clade – but diverging from the root, ~11 ka, and including other Filipinos and Malay within its diversity. Apart from haplogroups B4b1a2c and P9, haplogroups N11b (in negrito Mamanwa) and M80 (negrito Batak) are also found predominantly in the Philippine negrito and not elsewhere.

The genetic link between Malaysian negrito and other negrito populations in Southeast Asia therefore remains tenuous. My results appear consistent with Delfin and colleagues' (2011) view that there are no grounds for any inference of unique common ancestry. Rather there seems to be a common substrate for all of the populations throughout Southeast Asia. Indeed, haplogroups M42'M74 and R12'R21 both share a deep, ancient splits with Aboriginal Australian lineages (Figures 3.52 and 5.26 respectively), implying deep common links between the inhabitants of the Sunda and Sahul shelves. Convergent evolution, whilst a possibility, may in fact not be the only alternative hypothesis, as implied by Endicott (2013). My results support rather the mode of settlement captured in the single southern coastal route dispersal model (Macaulay *et al.*, 2005), with the implication that the various scattered negrito might themselves have remained relatively physically less changed from the early settlers, and their African ancestors, than other Eurasians, who adapted morphologically to



environments further north in Southeast Asia and then re-expanded southwards again from the Late Glacial onwards.

## 7.2 Aboriginal Malays (aka ‘Proto-Malays’)

As mentioned above, Aboriginal Malays share several intrusive lineages with the Semang and Senoi (such as N9a6, B5a1a, B4c2b and F1a1a) that appear to come into the Peninsula from northern MSEA during the late Pleistocene or early Holocene. However, the main intrusive lineage in Aboriginal Malays during the Holocene is R9b, which amount to almost a quarter of Temuan and Semelai lineages (Table 7.1). They also harbour indigenous Pleistocene founding mtDNA lineages (N21 and M22) distinct from those of the Semang and Senoi. N22 (or N22a on the whole-mtDNA tree) was also previously thought to be one of the Pleistocene founding lineages in Aboriginal Malays, but my study has shown that N22a expanded among Temuan Aboriginal Malays more recently (~2.5 ka; Figure 4.10). This might be associated with an arrival of Austronesian-speakers in Peninsular Malaysia from ISEA, but nevertheless the whole-mtDNA tree, although based on few samples, remains consistent with an indigenous origin within the Sunda region.

The source for R9b has also been controversial (Hill *et al.*, 2006). The deepest branches of the tree suggest a South Chinese or MSEA source, but the situation after ~20 ka remains unclear. The Aboriginal Malay lineages all cluster within R9b1a1a, dating to ~10 ka (Figure 5.30), which includes a further basal subclade restricted to the Peninsula (including Semang as well as Malay and Aboriginal Malay) and several in ISEA. An early Holocene dispersal into the Malay Peninsula/Sumatra seems the most likely explanation, but the nesting Philippine Mamanwa lineage makes a source in ISEA a possibility.

A dispersal south in the early Holocene from MSEA into the Peninsula might coincide with van Heekeren’s (1972) traditional view that the Hoabinhian originated in South China before spreading south to Peninsular Malaysia and North Sumatra around the terminal Pleistocene/Holocene period (Hill *et al.*, 2006). However, Mokhtar (2006) and subsequently Bulbeck (2011) argued that this is contradicted by archaeological evidence for the predominantly local development of the Malayan Hoabinhian evident at Bukit Bunuh, Malaysia. Bulbeck (2011) therefore suggests a terminal Pleistocene dispersal of R9b1a1a, ~10 ka, from northern Vietnam (then host to Hoabinhian/Bacsonian industries) or central

Thailand to central-western Sundaland and its southwards spread into southern Peninsular Malaysia as postglacial sea levels rose (see also Oppenheimer, 2011).

N21 and N22a both appear to be largely restricted to the Aboriginal Malay and are not found in the other *Orang Asli* groups (Figure 4.8), confirming the findings by Hill *et al.* (2007). Hill *et al.* (2007) also inferred from the HVS-I data that both N21 and N22 showed evidence of recent gene flow from ISEA. N21a1 is seen in both Aboriginal Malay and Peninsular Malay. However, in contrast to this, my resolved whole-mtDNA study shows that the deeper lineages of N21 appear to be restricted to MSEA/Sunda, suggesting a possible Late Glacial (~19 ka) MSEA origin with an early Holocene eastern Sunda spread. Similarly to R9b, an early Holocene expansion (~9 ka) from northern MSEA may have brought the N21 lineages into Peninsular Malaysia and the Aboriginal Malay, and also Indonesia, rather than an offshore source in ISEA. The Aboriginal Malays are nested within subclade N21a1a dating to ~4 ka, again indicating a certain degree of population subdivision.

The rarer haplogroup N22 shows a deep ancestry in Southeast Asia (~29 ka) with the Aboriginal Malay nested within the basal N22a (Figure 4.10), dating to ~2.5 ka, the long branch to this subclade suggesting substantial genetic drift. One interpretation might be that this suggests the arrival of Austronesian speakers from southern ISEA, east of Sumatra, in Peninsular Malaysia around 2.5 ka, in line with the standard model of Aboriginal Malays origins (Bellwood, 1997). This receives some support from the fact that its much more diverse sister clade, N22b, which dates to the LGM at ~25 ka, is shared across ISEA, and a third basal singleton lineage is seen in the Philippines. However, N22b shows a deep split dating to the LGM between the Malay Peninsula and ISEA lineages. It is therefore possible that N22 has an origin in glacial Sundaland, spreading across to the Philippines to the east. In any case, contrary to previous conclusions (Hill *et al.*, 2006), N22a does not seem to be one of the Pleistocene founding mtDNA lineages of the Aboriginal Malay.

The M7c3c starburst (Figure 3.15) offers a possible signal for the postulated Austronesian-speaking Neolithic dispersal from South China and Taiwan through ISEA into Peninsular Malaysia (Bellwood, 1997; Hill *et al.*, 2007). Given the confidence intervals on the age estimates, this remains possible. The estimated ages of M7c3c and M7c3c1 (~7.5 ka and ~6 ka) seem slightly too old for that archaeo-linguistic model, however, the phylogeography of the M7c3c1 clade might be more consistent with a Philippine/ISEA origin and a reverse migration to Taiwan. Archaeologically, red-slipped pottery is found at Gua

Kecil in Peninsular Malaysia, which is interpreted as reflecting late Neolithic or Early Metal Phase influence (Dunn, 1964). Red-slipped pottery is one of the suggested markers of Austronesian linguistic expansion (Bellwood, 1997; Bulbeck, 2008), and so its presence in Peninsular Malaysia could suggest an Austronesian association. Although red-slipped pottery appeared late in the Peninsular Malaysia sequence, and is found southwards in the vicinity of Gua Kecil (where Aboriginal Malays, both Southern Aslian and Malayic, are located), the M7c3c confidence intervals still fall within the window for this model with the introduction of haplogroups M7c3c as well as N22a, in Peninsular Malaysia, hence I cannot rule it out.

In conclusion, the whole-mtDNA analyses show that the Aboriginal Malay do not harbour any lineages that are clearly indigenous to the Peninsula, although some may be indigenous to MSEA/Sunda as a whole. As shown in Table 7.1, only haplogroups E and M7c3c indicate migrations from ISEA into the Aboriginal Malay, correlating with Bellwood's model but comprising only a small fraction of the total lineages (15/126 or 12%). The East Asian lineages of B4a1c and M7c1a are found in Semelai Aboriginal Malay but these singletons from Semelai could have been recent one-off occurrences in the Peninsular Malaysia samples, and may therefore be insignificant (discussed further in 7.3.3). The great majority of the Aboriginal Malay thus appears to derive their mtDNA ancestry from MSEA/Sunda, possibly including South China, a signal previously observed by Hill *et al.* (2006) and elaborated upon by Bulbeck (2008) and Oppenheimer (2011).

### 7.3 Modern Malay (aka 'Deutero-Malay')

In contrast to the *Orang Asli* (in particular the Aboriginal Malays – also known as 'Proto-Malays'), the mainstream Malay population of the Malay Peninsula (also known as 'Deutero-Malay') have been argued to be more recent immigrants from ISEA over the past 3–3.5 ka, bringing Austronesian languages and genes (ultimately originating in Taiwan in that model), *en bloc* to the Peninsula and largely replacing more ancient indigenes of the Peninsula (Bellwood, 1997). Predictions of this ISEA immigrant model can be tested on broad phylogeographic principles, by comparing the source of lineages found in the Malay populations surveyed here. Table 7.2 assigns Malay lineages (based on whole-mtDNA sequences) to one of four putative proximal source regions: East Asia, Island Southeast Asia/New Guinea, Mainland Southeast Asia/Sunda (i.e. ancient lineages indigenous to

Mainland Southeast Asia, including those found evenly in both in MSEA and western ISEA) and South Asia.

**Table 7.2 Assignment of Malay lineages to putative proximal source regions. Figures taken from Table 3.4.** (EA – East Asia, ISEA/NG – Island Southeast Asia/New Guinea, MSEA/SUN – Mainland Southeast Asia/Sunda, SAS – South Asia)

<b>Haplogroup</b>	<b>Putative proximal source</b>	<b>NEM</b>	<b>NWM</b>	<b>SEM</b>	<b>SWM</b>	<b>Total</b>	<b>%</b>
A	EA	0	0	1	0	1	0.3
B4a	EA	3	5	1	2	11	3.7
B4b1	EA	0	1	0	0	1	0.3
B5b	EA	0	3	1	0	4	1.3
C7a	EA	1	1	0	0	2	0.7
D4a3	EA	1	0	0	0	1	0.3
D5b	EA	0	0	1	0	1	0.3
F4b	EA	0	2	0	0	2	0.7
M7b	EA	1	2	0	7	10	3.4
M7b3	EA	0	0	2	0	2	0.7
N10	EA	2	3	0	0	5	1.7
R11b	EA	2	0	0	0	2	0.7
<b>Total East Asia</b>						<b>14.1</b>	
B4a1a	ISEA	4	3	1	0	8	2.7
B4c1b2	ISEA	3	4	4	10	21	7.1
E1a1a	ISEA	2	1	7	0	10	3.4
E1a2	ISEA	0	0	3	0	3	1.0
E1b	ISEA	3	3	1	0	7	2.4
E2a	ISEA	0	0	4	0	4	1.3
F3b	ISEA	1	0	0	0	1	0.3
M7c3c	ISEA	8	3	3	0	14	4.7
P1d	ISEA/NG	1	0	1	0	2	0.7
Q1	ISEA/NG	0	0	1	0	1	0.3
Q3	ISEA/NG	0	1	0	0	1	0.3
Y2a	ISEA	0	5	1	0	6	2.0
<b>Total ISEA/NG</b>						<b>26.3</b>	
B4c2	MSEA/SUN	1	4	0	0	5	1.7
B5a	MSEA/SUN	9	4	4	0	17	5.7
B6a1a	MSEA/SUN	0	1	1	3	5	1.7
F1a1	MSEA/SUN	2	2	0	0	4	1.3
F1a1a	MSEA/SUN	8	7	2	0	17	5.7
F1a3	MSEA/SUN	2	1	0	0	3	1.0
F1a4	MSEA/SUN	1	2	0	0	3	1.0
F1f	MSEA/SUN	6	2	0	1	9	3.0
F3a	MSEA/SUN	1	2	0	0	3	1.0
M*	MSEA/SUN	2	2	3	0	7	2.4
M12	MSEA/SUN	3	2	0	0	5	1.7

M13	MSEA/SUN	1	2	0	0	3	1.0
M17c	MSEA/SUN	5	1	0	0	6	2.0
M20	MSEA/SUN	6	2	3	0	11	3.7
M21a	MSEA/SUN	1	3	0	1	5	1.7
M21c	MSEA/SUN	2	1	0	0	3	1.0
M21d	MSEA/SUN	1	1	0	0	2	0.7
M22a	MSEA/SUN	0	2	0	0	2	0.7
M22b	MSEA/SUN	0	0	0	1	1	0.3
M26a	MSEA/SUN	3	0	1	0	4	1.3
M26b	MSEA/SUN	0	1	1	0	2	0.7
M47	MSEA/SUN	0	1	0	0	1	0.3
M50	MSEA/SUN	3	2	0	0	5	1.7
M51	MSEA/SUN	3	0	1	0	4	1.3
M71	MSEA/SUN	1	0	1	2	4	1.3
M72	MSEA/SUN	1	0	0	1	2	0.7
M73	MSEA/SUN	0	0	0	1	1	0.3
M74b	MSEA/SUN	0	3	0	0	3	1.0
M77	MSEA/SUN	1	0	0	0	1	0.3
N9a6	MSEA/SUN	1	0	0	1	2	0.7
N21	MSEA/SUN	2	1	0	1	4	1.3
N22	MSEA/SUN	1	0	0	1	2	0.7
N8	MSEA/SUN	0	0	1	0	1	0.3
R*	MSEA/SUN	0	5	0	0	5	1.7
R21	MSEA/SUN	1	0	0	0	1	0.3
R22	MSEA/SUN	2	1	5	0	8	2.7
R9b	MSEA/SUN	4	0	0	0	4	1.3
<b>Total MSEA/SUN</b>							<b>55.6</b>
M2b	SA	0	1	0	0	1	0.3
M30	SA	1	0	0	0	1	0.3
M32c	SA	0	0	1	0	1	0.3
M37	SA	0	0	0	1	1	0.3
M5	SA	0	1	0	0	1	0.3
R6a	SA	1	0	0	0	1	0.3
R7a	SA	0	0	0	1	1	0.3
U1a	SA	1	0	0	0	1	0.3
U2b1	SA	0	1	0	0	1	0.3
U7	SA	0	3	0	0	3	1.0
<b>Total South Asia</b>							<b>4.0</b>
<b>Total</b>		109	98	56	34	297	100.0

### 7.3.1 MSEA/Sunda haplogroups in the modern Malay

The ancient MSEA/Sunda haplogroups are found in more than half of the Malay samples at ~56% (Table 7.2). Many of the traces of ancient Sundaland are extremely rare in

the extant Malay populations where these lineages spread thinly across both MSEA and ISEA, and are putatively demonstrated by haplogroups M73'79, M47, M77, M71 and M72. Some examples from the results are highlighted here.

A particularly intriguing major clade is the putative M42'74, which appears to be of early Sunda origin, dating to ~60 ka, and has three basal lineages (Figure 3.52). M74 is seen across the Sunda region, with a Northwest Peninsular Malay found within subclade M74b1. There is also a basal paraphyletic lineage seen in a single Vietnamese individual, whereas M42 is seen in Aboriginal Australians, pointing to a deep ancestral connection between the first inhabitants of Sunda and Sahul that we would anticipate, given the southern route settlement model, but for which there is little direct evidence in the rest of the whole-mtDNA tree.

M13'46'61 shows deep links between South and Southeast Asia (Figures 3.41 and 3.42), which are not very common elsewhere in the mtDNA tree, although their significance is not yet clear. The *Orang Asli* and Malay are predominantly found within M13b1 dating to the LGM ~25 ka. Considering the basal pre-M13 lineage and the oldest subclade, M13b1 are found in Thailand and the Malay Peninsula respectively, whilst rare subclade M61b dates to the LGM and has been found in Vietnam and North Borneo, the evidence strongly indicates they are the relict descendants of the first settlers on the Sunda shelf. M61a shows a similar pattern to subclades M13a1 and M13b2, spreading northwards ~10 ka into China and Tibet, finally entering India more recently.

Haplogroup M1'20'51 also indicates a deep Sunda root at the time of primary settlement, found especially in MSEA/Peninsular Malaysia (Figures 3.47 and 3.48). The signal most probably indicates recent dispersals northwards into South China, and southwards in Sumatra, perhaps relating to the flooding of the Strait of Malacca in the early Holocene. M17 is an ancient Sunda haplogroup, dating to ~55 ka (Figure 3.29). The Semang Kensi are found within subclade M17a, dating to the LGM, where MSEA lineages predominate. The Malay, on the other hand, belong to haplogroup M17c, dating to ~47 ka. The M17 phylogeny overall is now widespread across Southeast Asia with long terminal branches indicating the preservation of ancient lineages in both Mainland and Island Southeast Asia. M50 is an ancient western Sunda lineage dating to ~55 ka with recent gene flow into ISEA. M12 dates to ~41 ka and predate the LGM. The overall distribution again suggests long-term ancestry in west Sunda/South China.

F1f shows a close connection between Malay and likely source populations in West Sunda and, especially, Sarawak in Borneo, who all speak the Austronesian Malayo-Polynesian languages. The results here therefore indicate that the F1f carriers in ISEA can trace their ancestry to western Sunda ~6 ka (Figure 5.33). F1a is found extensively across East Asia and Southeast Asia and dates to the LGM (Figures 5.33 and 5.34). F1a1a most likely originated in South China followed by rapid spread into MSEA in the early Holocene, perhaps coinciding with the coastal Neolithic dispersals. Interestingly, F1a1a appears to spread right down into Malaysia, rather than moving again in the mid- to late Holocene into the Peninsula. The latter movement was predicted by Bellwood's (1997) model for the spread of Austro-Asiatic speakers in the mid- to late Holocene into the Malay Peninsula in conjunction with south Thailand's Ban Kao culture. F1a1a then seems to have entered the ancestors of the Senoi and Aboriginal Malay from within the Peninsula ~4 ka, which would correspond with Bellwood's model (1993, 1997) for the appearance of Austro-Asiatic with respect to the timescale in which the Neolithic was brought into the Malaysian indigenous groups to create the ancestors of modern Senoi, but not the immediate source. F1a1a is well-represented at locations along the Mekong valley (Oppenheimer, 2011), and the early inhabitants of Laos appear to have been a suitable biological precursor for Thailand's late Neolithic to Bronze/Iron Age populations. Therefore, the evidence is at least consistent with Bulbeck's (2011: Figure 5) suggestions that F1a1a carriers took the Mekong route for early Austro-Asiatic dispersal and south into Peninsular Malaysia via the Isthmus of Kra (Sidwell and Blench, 2011). Once again, a greater diversity of F1a1 lineages seems to have been preserved in the modern Malay than in the *Orang Asli*. One small subclade is exclusively shared between a single modern Malay individual and an Aboriginal Malay.

There are single instance of Malay lineages within haplogroups F1a3a and F1a4a (Figure 5.33). Both of these haplogroups have been previously proposed as possible markers for Austronesian-speaker dispersals in Hill *et al.* (2007). The age of F1a3a, at ~10 ka, and its current distribution make this unlikely in this case. On the other hand, F1a4a dates to ~7 ka (with a 95% interval of ~2–11 ka), and considering it is dispersed thinly but widely in South China, Taiwan aboriginals and both MSEA and ISEA, F1a4a remains a possible candidate marker.

F3 has a wide but extremely rare distribution in China and MSEA, in particular for subclade F3a, while F3b1 is also found in South China, Taiwan, ISEA and modern Malay

(Figure 3.58). The Malay F3a carriers appear to have a northern MSEA/China origin with an upper bound on their dispersal south of ~12 ka. Taking another route, F3b1 probably has a source in South China and dispersed into Taiwan and ISEA by the mid-Holocene, with subsequent dispersal into ISEA and then Peninsular Malaysia ~4 ka.

R22 appears to be a surviving lineage from the initial founding of the Sunda shelf region, dating back to ~46 ka, and is widespread and present in mainstream indigenous groups throughout Mainland and Island SEA, including the Nicobars (Figure 5.27). These groups speak the Mon-Khmer language branch of the Austro-Asiatic family. This corresponds with Bellwood's (1997) and Higham's (2004) model for a Neolithic rice-farmer expansion in which Austro-Asiatic speakers arose in South China and, in parallel with Austronesian dispersals through Taiwan and ISEA, dispersed through MSEA, as far as the Nicobar Islands. However, we only have the HVS-I data of the Nicobars and no complete mtDNA sequences to examine this link more precisely.

N8 shows a deep Sunda distribution dating to the LGM, ~22 ka (Figure 4.12). The N8b2 lineages from Java, Indonesia and Thai appear to share a deep root centred on MSEA rather than in ISEA, although they were still connected as a single landmass before the first rapid sea-level rise ~14.5 ka. R11 is mainly restricted to China, and B6 is widely distributed in SEA. Both haplogroups R11 and B6 are rare and also present in MSEA (Northeast Malay) and they have quite dissimilar distributions. A modern Malay and Vietnamese nested within a Chinese clade R11b, dating to ~11 ka (Figure 5.24). B6 is largely found in MSEA/Peninsular Malaysia and dates to ~27 ka (Figure 5.25). Several Malay and Temuan have been found confined to B6a1a, dating to ~7 ka, most likely to follow the MSEA route into Peninsular Malaysia.

A very large fraction of modern Malay lineages therefore trace their ancestry to the ancient Sunda region. This implies very substantial recruitment locally from within the Peninsula by assimilation from *Orang Asli* groups as the modern Malay expanded, although it may also be the case that some arrived from ISEA within the last few thousand years, where they may have had an ancient Sunda origin in, for example, Sumatra or Borneo. Those with a more clearly offshore source are described next.



### 7.3.2 ISEA haplogroups in the modern Malay

Table 7.2 shows there are ~26% of the modern Malay lineages with characteristic ISEA origins in the Malay Peninsula (which include ~1.3% of Melanesian/New Guinean lineages P and Q; see below). Haplogroup E is an important component of mtDNA diversity in ISEA/Taiwan, which evolved *in situ* over the last 35 ka and dispersed extensively throughout ISEA in the early to mid-Holocene (Soares *et al.*, 2008). The Malay have low frequencies of E1a1 (Figure 3.23), E1a2 (Figure 3.25), E1b (along with Aboriginal Malays, Figure 3.23) and E2a (Figure 3.28), all dating to almost 8 ka. These lineages are likely to have arrived from ISEA in Peninsular Malaysia between ~4–8 ka. The genetic signature of haplogroup E may indicate the impact on coastal-dwelling populations of the rapid global warming that coincided with the three rapid sea-level rises, in particular its maritime orientation and the development of sailing technology (Oppenheimer, 1998; Solheim, 2006; Soares *et al.*, 2008), as seems to be also reflected in the BSPs for both the Malay and the *Orang Asli*.

B4a1a dates to ~8 ka and it is commonly seen throughout ISEA and in aboriginal Taiwanese, including Peninsular Malaysia and Thailand, as well as Papua New Guinea (Figure 5.5). Its starburst pattern points to a dramatic expansions across the region, centred on ISEA, in the early Holocene, similar to that seen in haplogroup E, i.e. too early for the out-of-Taiwan archaeo-linguistic hypothesis (Soares *et al.*, 2011).

Y2a1 does seem to fit better the out-of-Taiwan model for Neolithic Austronesian-speakers (Figure 4.6). Y is part of N9 and Y1 is restricted to northern East Asia, as with most of the N9 subclades, suggesting an origin in East Asia. Y2a dates to ~7 ka with a basal lineage found in the Taiwanese Saisiat ethnic group (Tabbada *et al.*, 2010), possibly supporting an origin amongst aboriginal Taiwanese. Y2a1 dates to ~5.5 ka, and has spread widely in ISEA, including both the Philippines (Gunnarsdóttir *et al.*, 2011a), and Indonesia (Tabbada *et al.*, 2010), and also to the Malay Peninsula. The two Northwest Peninsular Malays are of Aceh and Banjar ethnic groups that are commonly found in Northern Sumatra. Y2b is only represented here by a sample from Aichi in Japan, reported by Tanaka *et al.* (2004), again indicating an East Asian origin for Y2 and a recent dispersal into the Sunda region, possibly via Taiwan and the Philippines.

### **New Guinean Melanesian haplogroups in the Modern Malay**

New Guinean/Melanesian influences in Peninsular Malaysia are signalled by haplogroups Q and P, which amount to ~1.3% of the total. Haplogroup Q is an Oceanian haplogroup most commonly reported in Papua New Guinea and West Papua and at lower levels in Vanuatu, Polynesia and Micronesia (Figure 3.33). The HVS-I database shows it is also found at low levels in Banjarmasin and Kota Kinabalu of Borneo, Bali, Manado, Toraja, Ujung Padang and Sumba, and Peninsular Malaysia. There are only three Southeast Asian whole-mtDNA lineages (two Peninsular Malays and one Filipino) indicating haplogroup Q and clearly we need more extensive sampling within ISEA for a better phylogeographic picture. The most likely explanation is very recent Holocene gene flow from Near Oceania into ISEA as far as the Malay Peninsula, but more ancient minor dispersals into ISEA remain possible.

Haplogroup P is an indigenous Melanesian haplogroup. The modern Malay lineages are nested within subclade P1d1 (Figure 5.41), and similar to haplogroup Q, this could be the result of recent gene flow from Melanesia into ISEA. Intriguingly though, they form a subclade with a Late Glacial coalescence time; similarly, a Philippine subclade also has an ancient mid-Holocene date. In the case of the Malay subclade, the HVS-I network implies the existence of a related cluster that is indeed restricted to ISEA, indicating a small but significant and possibly ancient dispersal from New Guinea into ISEA, although it may only have reached the Malay Peninsula within the last few thousand years.

### **7.3.3 East Asian haplogroups in the modern Malay**

Table 7.2 shows that about ~14% of the modern Malay lineages have a likely source in East Asia. In haplogroup M7b, several instances of both Aboriginal Malays and modern Malays are scattered throughout the tree. The M7b Temuan and Malay lineages (of different subclades) appear to have maternal origin in East China during the late Pleistocene (~16 ka) and early-Holocene dispersals (~12-10 ka) south, rather than a Neolithic event coming from offshore. In a similar pattern, an Aboriginal Malay Seletar is found in G1c dating to ~13 ka, where G1 is largely found in East Asia and virtually absent in SEA and the Pacific. B4 likely originated in China and later dispersed both into northeast Asia and SEA, as seen also in B4b1a2 and minor other instances. Likewise, B4a1c shows a northern source in China dating to ~21 ka; an Aboriginal Malay Semelai and two Northern Peninsular Malay have been found

within this clade. Similar to the abovementioned sporadic instances, inferences about timing are difficult based on a few sequences, although they seem likely to be recent. Moreover, none of these subclades is identifiable by HVS data (and therefore it is not possible to show the detailed lineages in Table 7.1), and these occurrences could have been one-offs among the modern Malay samples.

A few modern Malay are found within several predominantly East Asian clades. Two are found in the Chinese and Japanese clades of D4a3 and D5b. Both Malay lineages may indicate intrusive dispersals from the north with an upper bound of ~7 ka and ~5 ka respectively – again, likely much more recent. C7a1 appears to have a far northern origin in north China dating to the Late Glacial ~16 ka, and although it is seen in Thailand and Peninsular Malaysia, they could be a result of a fairly recent event. A single Peninsular Malay lineage is found within subclade A5b sharing with lineages from China and Japan. Given that they are all vanishingly rare in Malaysia, they probably arrived quite recently; for example, there is a known history of immigrant influx from China and India since the 15<sup>th</sup>–17<sup>th</sup> centuries.

N10 shows a rare, deep and ancient root in East Asia (Figure 4.13). The Peninsular Malay lineage is found within a tiny clade N10a1 with an Indonesian lineage from South Borneo in Sunda, and the clade has a recent age ~2.5 ka, possibly due to genetic drift given then preceding long branch. Again, the timing is hard to infer based on so few sequences.

### **7.3.4 South Asian haplogroups in the modern Malay**

Haplogroup M2b is commonly found in the Dravidian speakers of central and south India, and in Korku, an Austro-Asiatic-speaking tribe of central India (Kumar *et al.*, 2008); unfortunately these complete sequences were not included in phylogenies for this study due to time constraints. The Malay lineage shares six M2b defining mutations (Figure 3.43), and M2b is estimated at ~13 ka by Soares *et al.* (2009), but my rho estimation is ~27 ka, implying a more ancient split. As with the Chinese lineages, the arrival time might well be very recent, within the last few centuries. Similarly, several Malay lineages are found within South Asian subclades M4'67 (including M37 and M30) and M5a, all dating to the Pleistocene (Figure 3.38). The Malay Peninsula has known historical contacts since around the 17<sup>th</sup> century, and these lineages again almost certainly arrived very recently.

On the other hand, the split time between M32a and M32c is at least consistent with the initial peopling process during the out of Africa dispersal. Interestingly, M32 was initially thought to be a specifically negrito Andamanese haplogroup that evolved within the Andaman Islands, but recent autosomal studies have suggested that the Andamanese are likely to have a Southeast Asian rather than indigenous or South Asian origin (Chaubey and Endicott, 2013). However, none of the negrito Semang lineages belongs to the haplogroup M31 and M32, lineages of the Andaman Islands, as discussed earlier.

Other smaller clades among the Malay haplogroups with Indian influences can be seen in haplogroups R6a (Figure 5.43) and R7a, both showing deep roots in India. R6a dates to ~52 ka and singleton Malay and Thai lineages are nested within subclade R6a1b (~37 ka). A Southwest Peninsular Malay lineage is seen within the derived Indian subclade R7a1a, dating to ~4.4 ka (Figure 5.44). Given the known historical connections between India and the Malay Peninsula, this is likely to have arrived recently in Malaysia.

On the other hand, a novel singleton branch found in this study in North Borneo very deep within R7 (Figure 5.44), suggests a potentially much earlier dispersal between South and Southeast Asia, possibly dating back as early as 55 ka, in which case it may date to the time of the southern-route dispersal from South to Southeast Asia itself.

## 7.4 Conclusions

This characterisation of the whole-mtDNA of the *Orang Asli* and modern Malay populations of Peninsular Malaysia shows the presence of both indigenous clades and genetic influences from outside the Peninsula. In the case of all three groups of *Orang Asli*, especially the Semang, it shows that they have experienced high levels of genetic drift resulting in a small number of sequence types elevated to very high frequencies, as demonstrated by haplogroups M21a1b, M22a2, N21a1a, N22a, N9a6 and R21. However, paradoxically, given their traditional status as relative newcomers to the Peninsula, much higher levels of clearly indigenous diversity remain among the modern Malay.

All three *Orang Asli* groups appear to mainly descend from indigenous Pleistocene populations and to have received substantial multiple maternal lineage influences from northern Indo-China during the Holocene, before the spread of rice agriculture. It is apparent that the traditional models are too simple to explain the complexity of population history in

Peninsular Malaysia. The assumption of unchanged relicts of earlier population waves seems completely unfounded. All three *Orang Asli* groups have local roots that trace back to ~50 ka, and all have been affected to a greater or lesser extent by subsequent migrations to Peninsular Malaysia. The Semang and Senoi show much lesser haplogroup diversity compared to the Aboriginal Malays. The latter show some connections with ISEA and even East Asia, but they harbour haplogroups that are either novel or rare elsewhere, a diverse composition matching the mtDNA gene pool of modern Malay in Peninsular Malaysia. While some of the ancestors of the Aboriginal Malays could have taken part in the colonisation of the Indo-Malaysian Archipelago during the past 3–3.5 ka, it is apparent that these would represent only a small portion of the maternal lineages. Despite the large cultural impact of the Austronesian expansion, there are only minor impacts from the Neolithic incursions on both maternal and paternal gene pool in the Peninsula.

The Aboriginal Malays have some indigenous ancestry that is as deep as that of the Semang and Senoi in Peninsular Malaysia. They exhibit less extreme patterns of genetic drift than the Semang, perhaps reflecting their larger population size (~40,000) compared with the Semang (~2000) (Senoi ~49,000 in year 2000; Benjamin, 2002b). The Senoi, although now the largest group in census size, appear to have undergone more recent drift than the Aboriginal Malays. This may be due to the initial processes of ethnogenesis or subsequent founder effects, such as the proposed expansion of the Temiar eastwards in recent times (Benjamin, 2002a).

The MSY (male-specific region of the Y chromosome) studies of the Austronesian-populations in ISEA (as well as Oceania) showed the majority of their paternal heritage tracing to the first Pleistocene settlers, with a smaller fraction tracing to more-recent immigration from northern MSEA (Capelli *et al.*, 2001; Karafet *et al.*, 2010; He *et al.*, 2012). A similar signal was found by Simonson *et al.* (2011) in the Austronesian-speaking Iban of Sarawak, except there is no northern influence such as from Taiwan found in the Iban. The autosomal SNP markers indicated a majority of the populations in the vicinity have a Pleistocene source in ISEA (Abdulla *et al.*, 2009). Wong *et al.* (2013) found rare low-frequency variants in the Singaporean Malay that were not found before, a finding perhaps implying, in similar fashion to the findings of the present study, that the range of ancient Sundaland mtDNA lineage diversity is preserved better in the modern Malay than in the *Orang Asli*. More autosomal data from additional populations combined with demographic

modelling are required to sort out the relative roles of residence pattern, society structure, amount of admixture, and subsequent role of migration and drift in shaping the gene pool of the region (Friedlaender *et al.*, 2008; Kayser *et al.*, 2008).

The study of modern Malay complete-mtDNA sequences has shown a high diversity of mtDNA lineages in Peninsular Malaysia and demonstrated the maternal contribution of several distinct regions to the history and ancestry of the Malay Peninsula: perhaps even more than most populations, the Malay are a composite with many different ancestries. The high level of maternal diversity indicates that the Malay have remained at a much larger effective population size over long periods of time, and thus have not been as susceptible to genetic drift as the *Orang Asli* groups. The Malay lineages have specifically shown many ancestral indigenous lineages related to those of the *Orang Asli*, which have evidently been lost from the *Orang Asli* but survived at a significant level in the Malay populations – indeed, forming a majority of the maternal lineages of the Malay. This implies that the most powerful approach to phylogeographic analysis of modern human populations, where possible, is to combine the use of “relict” and mainstream population lineages for a full phylogeographic picture, and not just to focus on the former.

Apart from the many indigenous lineages that can be traced to the first settlers in Sundaland, which form a majority of the Malay maternal ancestry, the Malay populations appear to have had extensive maternal genetic influences from both East Asia and ISEA, as far east as Near Oceania, as well as (to a lesser extent) the Indian Subcontinent, at different periods of time. The conventional Bellwood model (1993, 1997) suggested that most of the Malay ancestors would have come into the Malay Peninsula over the past 3,000 years as intrusive migrations from ISEA. My data have indicated that they form a diverse, composite group, rather different from the populations of ISEA, with lineages from ISEA making up little more than a quarter of Malay maternal ancestry. This includes extensive sharing with Sumatra and Borneo, but further analysis will be necessary to tease out the timing and direction of migrations at this level.

These signals reflect a complex and dynamic demographic history among the populations in Peninsular Malaysia as a result of climate change which is unique to Sundaland. The Bayesian Skyline Plots (BSPs) suggest a decline in the effective population size after the LGM and a major crash after ~11 ka – likely due to the devastating effect of sea-level rises – followed by rapid recovery ~7 ka, as some populations re-adapted to coastal

living and expanded along the extended coastlines that had become available. This was previously predicted for ISEA from the phylogeography of haplogroup E, but the fact that it is echoed in the BSP for the *Orang Asli* indicates that it affected MSEA as well as ISEA populations, as would be expected from their common ancestry in the Pleistocene Sunda population.

The high resolution of complete-mtDNA sequencing of samples from Peninsular Malaysia and the phylogeographic analysis, in this study and others, have been crucial to this improved understanding at a fine level of detail the genealogical relations and ages of lineages both within these populations and with other groups throughout Southeast Asia, Taiwan and East Asia, and even with South Asia and the western Pacific, in order to untangle their complex history of migration and settlement. Clearly, the history of Peninsular Malaysia is much too complex to be explained by any simple model. The huge reservoir of variation revealed by this study suggests that simple migration and replacement models are far too crude to explain the data. Some migratory events have clearly taken place but not from only one direction, and in each case since the first settlement they can be seen as having enriched the variation already present.

## References

- Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C., Chaurasia, A., Chen, C. H., Chen, J., Chen, Y. T., Chu, J., Cutiongco-de la Paz, E. M., De Ungria, M. C., Delfin, F. C., Edo, J., Fuchareon, S., Ghang, H., Gojobori, T., Han, J., Ho, S. F., Hoh, B. P., Huang, W., Inoko, H., Jha, P., Jinam, T. A., Jin, L., Jung, J., Kangwanpong, D., Kampuansai, J., Kennedy, G. C., Khurana, P., Kim, H. L., Kim, K., Kim, S., Kim, W. Y., Kimm, K., Kimura, R., Koike, T., Kulawonganunchai, S., Kumar, V., Lai, P. S., Lee, J. Y., Lee, S., Liu, E. T., Majumder, P. P., Mandapati, K. K., Marzuki, S., Mitchell, W., Mukerji, M., Naritomi, K., Ngamphiw, C., Niikawa, N., Nishida, N., Oh, B., Oh, S., Ohashi, J., Oka, A., Ong, R., Padilla, C. D., Palittapongarnpim, P., Perdigon, H. B., Phipps, M. E., Png, E., Sakaki, Y., Salvador, J. M., Sandraling, Y., Scaria, V., Seielstad, M., Sidek, M. R., Sinha, A., Srikummool, M., Sudoyo, H., Sugano, S., Suryadi, H., Suzuki, Y., Tabbada, K. A., Tan, A., Tokunaga, K., Tongsimma, S., Villamor, L. P., Wang, E., Wang, Y., Wang, H., Wu, J. Y., Xiao, H., Xu, S., Yang, J. O., Shugart, Y. Y., Yoo, H. S., Yuan, W., Zhao, G. & Zilfalil, B. A. 2009. Mapping human genetic diversity in Asia. *Science*, 326, 1541-5.
- Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q.-P., Woodward, S. R., Salas, A., Torroni, A. & Bandelt, H.-J. 2008. The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies. *PLoS One*, 3, e1764.
- Adelaar, K. A. 1994. The classification of the Tamanic languages. In: Dutton, T. & Tryon, D. (eds.) *Contact-induced change in Austronesian languages*. Berlin: Mouton de Gruyter.
- Adelaar, K. A. 2006. Borneo as a Cross-Roads for Comparative Austronesian Linguistics. In: Bellwood, P., Fox, J. J. & Tryon, D. (eds.) *The Austronesians: Historical and Comparative Perspectives*. Australian National University: ANU E Press.
- Adi, H. T. 2000. *Archaeological Excavations in Ulu Kelantan, Peninsular Malaysia*. PhD, Australian National University.
- Aitken, M. J. 1998. *An Introduction to Optical Dating: The Dating of Quaternary Sediments by the Use of Photon-Stimulated Luminescence*, Oxford, Oxford University Press.
- Allen, J., Gosden, C., Jones, R. & White, J. P. 1988. Pleistocene dates for the human occupation of New Ireland, northern Melanesia. *Nature*, 331, 707-9.
- Ambrose, S. H. 1998. Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. *J Hum Evol*, 34, 623-651.
- Anderson, D. D. 1990. *A Lang Rongrien Rockshelter: A Pleistocene, Early Holocene Archaeological Site from Krabi, Southwestern Thailand*, University Museum, University of Pennsylvania.
- Anderson, D. D. 1997. Cave archaeology in Southeast Asia. *Geoarchaeology*, 12, 607-638.
- Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. & Young, I. G. 1981. Sequence and organization of the human mitochondrial genome. *Nature*, 290, 457-65.
- Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 23, 147.



- Appenzeller, T. 2012. Human migrations: Eastern odyssey. *Nature*, 485, 24-6.
- Arnason, U., Xu, X. & Gullberg, A. 1996. Comparison between the complete mitochondrial DNA sequences of Homo and the common chimpanzee based on nonchimeric sequences. *J Mol Evol*, 42, 145-52.
- Arredi, B., Poloni, E. S., Paracchini, S., Zerjal, T., Fathallah, D. M., Makrelouf, M., Pascali, V. L., Novelletto, A. & Tyler-Smith, C. 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet*, 75, 338-45.
- Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A. & Saunders, N. C. 1987. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, 18, 489-522.
- Awadalla, P., Eyre-Walker, A. & Smith, J. M. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*, 286, 2524-5.
- Ayala, F. J. 1995. The myth of Eve: molecular biology and human origins. *Science*, 270, 1930-6.
- Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., Chen, K. H. & Wallace, D. C. 1992. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient mongoloid migrations. *Genetics*, 130, 139-52.
- Bandelt, H. J. & Dur, A. 2007. Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. *Mol Phylogenet Evol*, 42, 256-71.
- Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. 1995. Mitochondrial portraits of human populations using median networks. *Genetics*, 141, 743-53.
- Bandelt, H. J., Kong, Q. P., Parson, W. & Salas, A. 2005. More evidence for non-maternal inheritance of mitochondrial DNA? *J Med Genet*, 42, 957-60.
- Bandelt, H. J., Lahermo, P., Richards, M. & Macaulay, V. 2001. Detecting errors in mtDNA data by phylogenetic analysis. *Int J Legal Med*, 115, 64-9.
- Bandelt, H. J., Quintana-Murci, L., Salas, A. & Macaulay, V. 2002. The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet*, 71, 1150-60.
- Bandelt, H. J., Salas, A. & Lutz-Bonengel, S. 2004. Artificial recombination in forensic mtDNA population databases. *Int J Legal Med*, 118, 267-73.
- Banks, W. E., d'Errico, F., Peterson, A. T., Vanhaeren, M., Kageyama, M., Sepulchre, P., Ramstein, G., Jost, A. & Lunt, D. 2008. Human ecological niches and ranges during the LGM in Europe derived from an application of eco-cultural niche modeling. *Journal of Archaeological Science*, 35, 481-491.
- Barker, G. 2005. The archaeology of foraging and farming at Niah Cave, Sarawak. *Asian Perspectives*, 44, 90-106.
- Barker, G., Reynolds, T. & Gilbertson, D. 2005. The Human Use of Caves in Peninsular and Island Southeast Asia: Research Themes. *Asian Perspectives*, 44, 1-15.
- Beavitt, P., Kurui, E. & Thompson, G. 1996. Confirmation of an early date for the presence of rice in Borneo: evidence for possible Bidayuh/Asian links. *Borneo research bulletin*, 27, 29-38.
- Behar, D. M., Hammer, M. F., Garrigan, D., Villems, R., Bonne-Tamir, B., Richards, M., Gurwitz, D., Rosengarten, D., Kaplan, M., Della Pergola, S., Quintana-Murci, L. & Skorecki, K. 2004. MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. *Eur J Hum Genet*, 12, 355-64.
- Behar, D. M., van Oven, M., Rosset, S., Metspalu, M. t., Loogvääli, E.-L., Silva, N. M., Kivisild, T., Torroni, A. & Villems, R. 2012. A Copernican Reassessment of the Human Mitochondrial DNA Tree from its Root. *Am J Hum Genet*, 90, 675-684.

- Behar, D. M., VILLEMS, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., Bertranpetit, J., Quintana-Murci, L., Tyler-Smith, C., Wells, R. S. & Rosset, S. 2008. The dawn of human matrilineal diversity. *Am J Hum Genet*, 82, 1130-40.
- Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M. F., Skorecki, K. & VILLEMS, R. 2010. The genome-wide structure of the Jewish people. *Nature*, 466, 238-42.
- Bellwood, P. 1987. *The Polynesians: Prehistory of an Island People*, London, Thames and Hudson.
- Bellwood, P. 1988a. *Archaeological Research in South-eastern Sabah*, Kota Kinabalu, Sabah Museum Monograph.
- Bellwood, P. 1989. Archaeological investigation at Bukit Tengkorak and Segarong, southeastern Sabah. *Bulletin of the Indo-Pacific Prehistory Association*, 9, 122-162.
- Bellwood, P. 1990. From Late Pleistocene to early Holocene in Sundaland. In: Gamble, C. & Soffer, O. (eds.) *The World at 18,000 B.P., Vol. 2: Low Latitudes*. London: Unwin Hyman.
- Bellwood, P. 1993. Cultural and biological differentiation in Peninsular Malaysia: The last 10,000 years. *Asian Perspectives*, 32, 37-60.
- Bellwood, P. 1994. *Southeast Asia before History*, Singapore, Cambridge University Press.
- Bellwood, P. 1997. *Prehistory of the Indo-Malaysian Archipelago*, Honolulu, Hawaii, University of Hawai'i Press.
- Bellwood, P. 2001. Early agriculturalist population diasporas? Farming, languages, and genes. *Annual Review of Anthropology*, 30, 181-207.
- Bellwood, P. 2005a. *First Farmers: the origins of agricultural societies*, Malden, MA, Blackwell Publisher.
- Bellwood, P. 2005b. *Examining the Farming/Language Dispersal Hypothesis in the East Asian Context*, Oxon, RoutledgeCurzon.
- Bellwood, P. 2011. Holocene Population History in the Pacific Region as a Model for Worldwide Food Producer Dispersals. *Current Anthropology*, 52, S363-S378.
- Bellwood, P. & Dizon, E. 2005. The Batanes Archaeological Project and the "Out of Taiwan" Hypothesis for Austronesian Dispersal. *Journal of Austronesian Studies*, 1, 1-33.
- Bellwood, P. & Dizon, E. 2008. Austronesian cultural origins: Out of Taiwan, via the Batanes Islands, and onwards to western Polynesia. In: Sanchez-Mazas, A., Blench, R., Ross, M. D., Peiros, I. & Lin, M. (eds.) *Past Human Migrations in East Asia: Matching Archaeology, Linguistics and Genetics*. London: Routledge.
- Bellwood, P., Fox, J. J. & Tryon, D. 2006. *The Austronesians: Historical and Comparative Perspectives*, Canberra, ANU E Press.
- Bellwood, P. & Renfrew, C. 2003. *Examining the Farming/Languages Dispersal Hypothesis*, Cambridge, McDonald Institute for Archaeological Research.
- Bellwood, P., Stevenson, J., Dizon, E., Mijares, A., Lacsina, G. & Robles, E. 2008. Where are the Neolithic Landscapes of Ilocos Norte? *Hukay*, 13, 25-38.
- Bendall, K. E. & Sykes, B. C. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet*, 57, 248-56.
- Benjamin, G. 1979. Indigenous religious systems of the Malay Peninsula. In: Becker, A. & Yengoyan, A. (eds.) *The imagination of reality: essays in Southeast Asian coherence systems*. Norwood: Ablex Publishing.

- Benjamin, G. 1985. In the long term: Three themes in Malaysian cultural ecology. *In: Hutterer, K. L., Rambo, A. T. & Lovelace, G. (eds.) Cultural values and human ecology in Southeast Asia*. Ann Arbor (MI): Michigan University Press.
- Benjamin, G. 1986. Between Isthmus and Islands: reflections on Malayan Palaeo-sociology. *Working Papers Issue 71*. Singapore: Department of Sociology, National University of Singapore.
- Benjamin, G. 1997. Issues in the Ethnohistory of Pahang. *In: Nik Hassan Shuhaimi, N. A. R., Mohamed Mokhtar, A. B., Ahmad, H. K. & Jazamuddin, B. (eds.) Pembangunan Arkeologi Pelancongan Negeri Pahang*. Pahang: Lembaga Muzium Negeri Pahang.
- Benjamin, G. 2002a. On being tribal in the Malay world. *In: Benjamin, G. & Chou, C. (eds.) Tribal communities in the Malay world: historical, cultural and social perspectives*. Singapore: Institute for Southeast Asian Studies.
- Benjamin, G. 2002b. Orang Asli languages: from heritage to death? *In: Rashid, R. & Wazir, J. K. (eds.) Minority cultures of Peninsular Malaysia: survivals of indigenous heritage*. Penang (Malaysia): AKASS (Malaysian Academy of Social Science).
- Beyin, A. 2011. Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate. *Int J Evol Biol*, 2011, 615094.
- Bird, M. I., Taylor, D. & Hunt, C. 2005. Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? *Quaternary Science Reviews*, 24, 2228-2242.
- Birdsell, J. B. 1993. *Microevolutionary Patterns in Aboriginal Australia: A Gradient Analysis of Clines*, Oxford Oxford University Press.
- Black, M. L., Dufall, K., Wise, C., Sullivan, S. & Bittles, A. H. 2006. Genetic ancestries in northwest Cambodia. *Ann Hum Biol*, 33, 620-7.
- Blanchon, P. & Shaw, J. 1995. Reef drowning during the last deglaciation: Evidence for catastrophic sea-level rise and ice-sheet collapse. *Geology*, 23, 4-8.
- Blevins, J. 2007. A Long Lost Sister of Proto-Austronesian? Proto-Ongan, Mother of Jarawa and Onge of the Andaman Islands. *Oceanic Linguistics* 46, 154-198.
- Blust, R. 1995. The prehistory of the Austronesian-speaking peoples: A view from language. *Journal of World Prehistory*, 9, 453-510.
- Blust, R. 1996. Austronesian Culture History: The Window of Language. *Transactions of the American Philosophical Society, New Series*, 86.
- Blust, R. 1999. Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. *In: Zeitoun, E. & Li, P. J.-K. (eds.) Selected Papers From the 8th International Conference on Austronesian Linguistics*. Taipei, Taiwan: Academia Sinica.
- Blust, R. 2013. Terror from the sky: unconventional linguistic clues to the negrito past. *Hum Biol*, 85, 401-16.
- Bodner, M., Zimmermann, B., Rock, A., Kloss-Brandstatter, A., Horst, D., Horst, B., Sengchanh, S., Sanguansermsri, T., Horst, J., Kramer, T., Schneider, P. M. & Parson, W. 2011. Southeast Asian diversity: first insights into the complex mtDNA structure of Laos. *BMC Evol Biol*, 11, 49.
- Bolnick, D. A., Bolnick, D. I. & Smith, D. G. 2006. Asymmetric male and female genetic histories among Native Americans from Eastern North America. *Mol Biol Evol*, 23, 2161-74.
- Briggs, A. W., Good, J. M., Green, R. E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V. B., Golovanova, L. V., de la Rasilla, M., Fortea, J., Rosas, A. & Paabo, S. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, 325, 318-21.

- Bromham, L. & Penny, D. 2003. The modern molecular clock. *Nat Rev Genet*, 4, 216-24.
- Brown, D. T., Samuels, D. C., Michael, E. M., Turnbull, D. M. & Chinnery, P. F. 2001. Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *Am J Hum Genet*, 68, 533-6.
- Brown, P., Sutikna, T., Morwood, M. J., Soejono, R. P., Jatmiko, Saptomo, E. W. & Due, R. A. 2004. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature*, 431, 1055-61.
- Brown, W. M., George, M., Jr. & Wilson, A. C. 1979. Rapid evolution of animal mitochondrial DNA. *Proc Nat Acad Sci USA*, 76, 1967-71.
- Bulbeck, D. 1996. Holocene Biological Evolution of the Malay Peninsula Aborigines (Orang Asli). *Perspectives in Human Biology*. World Scientific.
- Bulbeck, D. 2000. Dental morphology at Gua Cha, West Malaysia, and the implications for "Sundadonty". *Bulletin of the Indo-Pacific Prehistory Association*, 19, 17-41.
- Bulbeck, D. 2003. Hunter-Gatherer Occupation of the Malay Peninsula from the Ice Age to the Iron Age. In: Mercader, J. (ed.) *Under the canopy: The archaeology of tropical rain forests*. Piscataway, NJ: Rutgers University Press.
- Bulbeck, D. 2004a. Indigenous traditions and exogenous influences in the early history of Peninsular Malaysia. In: Glover, I. & Bellwood, P. (eds.) *Southeast Asia from prehistory to history*. London and New York: RoutledgeCurzon.
- Bulbeck, D. 2008. An Integrated Perspective on the Austronesian Diaspora: The Switch from Cereal Agriculture to Maritime Foraging in the Colonisation of Island Southeast Asia. *Australian Archaeology*, 67, 31-52.
- Bulbeck, D. 2011. Biological and cultural evolution in the population and culture history of *Homo sapiens* in Malaya. In: Enfield, N. J. (ed.) *Dynamics of human diversity: the case of mainland Southeast Asia*. Canberra Australia: Pacific Linguistics.
- Bulbeck, D. 2013. Craniodental affinities of Southeast Asia's "negritos" and the concordance with their genetic affinities. *Hum Biol*, 85, 95-133.
- Burenhult, N. 2001. Jahai phonology: A preliminary survey. *The Mon-Khmer Studies Journal*, 31, 29-45.
- Burenhult, N., Kruspe, N. & Dunn, M. 2011. Language history and culture groups among Austroasiatic-speaking foragers of the Malay Peninsula. In: Enfield, N. J. (ed.) *Dynamics of human diversity: the case of mainland Southeast Asia*. Canberra, Australia: Pacific Linguistics.
- Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., Wei, L., Wang, C., Li, S., Huang, X., Jin, L. & Li, H. 2011. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. *PLoS One*, 6, e24282.
- Cann, R. L., Stoneking, M. & Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature*, 325, 31-6.
- Capelli, C., Wilson, J. F., Richards, M., Stumpf, M. P., Gratrix, F., Oppenheimer, S., Underhill, P., Pascali, V. L., Ko, T. M. & Goldstein, D. B. 2001. A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am J Hum Genet*, 68, 432-43.
- Carey, I. 1976. *Orang Asli: The Aboriginal Tribes of Peninsula Malaysia*, Kuala Lumpur, Oxford University Press.
- Carvajal-Carmona, L. G., Soto, I. D., Pineda, N., Ortiz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V. M., Bedoya, G. & Ruiz-Linares, A. 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet*, 67, 1287-95.

- Cavalli-Sforza, L. L. & Feldman, M. W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat Genet*, 33 Suppl, 266-75.
- Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994. *The history and geography of human genes*, Princeton, N.J., Princeton University Press.
- Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S., Bandopadhyay, S. S., Barua, P., Barik, S. S., Basu, D., Kiran, U., Gangopadhyay, P., Sahani, R., Prasad, B. V., Gangopadhyay, S., Lakshmi, G. R., Ravuri, R. R., Padmaja, K., Venugopal, P. N., Sharma, M. B. & Rao, V. R. 2009. Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. *PLoS One*, 4, e7447.
- Chaubey, G. & Endicott, P. 2013. The Andaman Islanders in a regional genetic context: reexamining the evidence for an early peopling of the archipelago from South Asia. *Hum Biol*, 85, 153-72.
- Chaubey, G., Karmin, M., Metspalu, E., Metspalu, M., Selvi-Rani, D., Singh, V. K., Parik, J., Solnik, A., Naidu, B. P., Kumar, A., Adarsh, N., Mallick, C. B., Trivedi, B., Prakash, S., Reddy, R., Shukla, P., Bhagat, S., Verma, S., Vasnik, S., Khan, I., Barwa, A., Sahoo, D., Sharma, A., Rashid, M., Chandra, V., Reddy, A. G., Torroni, A., Foley, R. A., Thangaraj, K., Singh, L., Kivisild, T. & Villems, R. 2008. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. *BMC Evol Biol*, 8, 227.
- Chen, T. & Zhang, Y. 1991. Palaeolithic chronology and possible coexistence of *Homo erectus* and *Homo sapiens* in China. *World Archaeol*, 23, 147-54.
- Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S. & Wallace, D. C. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet*, 57, 133-49.
- Chia, S. 2003. The prehistory of Bukit Tengkorak, Sabah, Malaysia. *Journal of Southeast Asian Archaeology*, 21, 146-159.
- Chinnery, P. 2006. Mitochondrial DNA in *Homo Sapiens*. In: Bandelt, H.-J., Macaulay, V. & Richards, M. (eds.) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer Berlin Heidelberg.
- Clark, J. D., Beyene, Y., WoldeGabriel, G., Hart, W. K., Renne, P. R., Gilbert, H., Defleur, A., Suwa, G., Katoh, S., Ludwig, K. R., Boissierie, J. R., Asfaw, B. & White, T. D. 2003. Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*, 423, 747-52.
- Clark, P. U. & Mix, A. C. 2002. Ice sheets and sea level of the Last Glacial Maximum. *Quaternary Science Reviews*, 21, 1-7.
- Clarkson, C., Petraglia, M., Korisettar, R., Haslam, M., Boivin, N., Crowther, A., Ditchfield, P., Fuller, D., Miracle, P., Harris, C., Connell, K., James, H. & Koshy, J. 2009. The oldest and longest enduring microlithic sequence in India: 35 000 years of modern human occupation and change at the Jwalapuram Locality 9 rockshelter. *Antiquity*, 83, 326-348.
- Cole, F. C. 1945. *The peoples of Malaysia*, Princeton (New Jersey), D. van Nostrand Co.
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E., Martinez-Arias, R., Clarimon, J., Facchini, F., Fiori, G., Luiselli, D., Pettener, D. & Bertranpetit, J. 1998. Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am J Hum Genet*, 63, 1824-38.
- Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F. & Bertranpetit, J. 2004. Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet*, 12, 495-504.

- Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E. & Awadalla, P. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet*, 43, 712-4.
- Consortium, I. H. 2003. The International HapMap Project. *Nature*, 426, 789-96.
- Costa, M. D., Cherni, L., Fernandes, V., Freitas, F., Ammar El Gaaied, A. B. & Pereira, L. 2009. Data from complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association and the "golden mean" to longevity. *Mech Ageing Dev*, 130, 222-6.
- Cox, M. P., Karafet, T. M., Lansing, J. S., Sudoyo, H. & Hammer, M. F. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc Biol Sci*, 277, 1589-96.
- Cree, L. M., Samuels, D. C., de Sousa Lopes, S. C., Rajasimha, H. K., Wonnapijit, P., Mann, J. R., Dahl, H. H. & Chinnery, P. F. 2008. A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. *Nat Genet*, 40, 249-54.
- Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., Wallace, D. C., Oefner, P. J., Torroni, A., Cavalli-Sforza, L. L., Scozzari, R. & Underhill, P. A. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*, 70, 1197-214.
- Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. & Scozzari, R. 2011. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am J Hum Genet*, 88, 814-8.
- Cruz, S. D., Parone, P. A. & Martinou, J. C. 2005. Building the mitochondrial proteome. *Expert Reviews of Proteomics*, 2, 541-551.
- Dancause, K. N., Chan, C. W., Arunotai, N. H. & Lum, J. K. 2009. Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. *J Hum Genet*, 54, 86-93.
- de Knijff, P. 2000. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet*, 67, 1055-61.
- Deacon, H. J. 1992. Southern Africa and modern human origins. *Phil Trans R Soc Lond B*, 337, 177-183.
- Deacon, H. J. & Geleijnse, V. B. 1988. The Stratigraphy and Sedimentology of the Main Site Sequence, Klasies River, South Africa. *The South African Archaeological Bulletin*, 43, 5-14.
- Delfin, F., Salvador, J. M., Calacal, G. C., Perdigon, H. B., Tabbada, K. A., Villamor, L. P., Halos, S. C., Gunnarsdottir, E., Myles, S., Hughes, D. A., Xu, S., Jin, L., Lao, O., Kayser, M., Hurles, M. E., Stoneking, M. & De Ungria, M. C. 2011. The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. *Eur J Hum Genet*, 19, 224-30.
- Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., Dorzhu, C., Luzina, F., Lee, H. K., Vanecsek, T., Villems, R. & Zakharov, I. 2007. Phylogeographic analysis of mitochondrial DNA in northern Asian populations. *Am J Hum Genet*, 81, 1025-41.

- Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G. & Capelli, C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol*, 21, 1673-82.
- Diamond, J. & Bellwood, P. 2003. Farmers and Their Languages: The First Expansions. *Science*, 300, 597-603.
- Diamond, J. M. 1988. Express train to Polynesia. *Nature*, 336, 307-308.
- Donohue, M. & Denham, T. 2010. Farming and Language in Island Southeast Asia: Reframing Austronesian History. *Current Anthropology*, 51, 223-256.
- Donohue, M. & Grimes, C. E. 2008. Yet more on the position of the languages of eastern Indonesia and East Timor. *Oceanic Linguistics*, 47, 114-158.
- Drummond, A. & Rambaut, A. 2009. Bayesian evolutionary analysis by sampling trees. In: Salemi, M., Vandame, A-M., Lemey, P. (ed.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. Cambridge: Cambridge University Press.
- Drummond, A. J. & Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7, 214.
- Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*, 22, 1185-92.
- Dubut, V., Cartault, F., Payet, C., Thionville, M.-D. & Murail, P. 2009. Complete Mitochondrial Sequences for Haplogroups M23 and M46: Insights into the Asian Ancestry of the Malagasy Population. *Hum Bio*, 81, 495-500.
- Dunn, F. 1964. Excavations at Gua Kechil, Pahang. *Journal of the Malaysian Branch of the Royal Asiatic Society*, 37, 87-124.
- Dyen, I. 1965. *A lexicostatistical classification of the Austronesian languages*.
- Eder, J. F. 1987. *On the Road to Tribal Extinction: Depopulation, Deculturation, and Adaptive Well-Being Among the Batak of the Philippines*, Berkeley, University of California Press.
- Eltsov, N. P. & Volodko, N. V. 2011. *MtPhyl: Software tool for human mtDNA analysis and phylogeny reconstruction* [Online]. Available: <https://sites.google.com/site/mtphyl/downloads> [Accessed 20/01/2012 2012].
- Endicott, P. 2013. Introduction: revisiting the "negrito" hypothesis: a transdisciplinary approach to human prehistory in southeast Asia. *Hum Biol*, 85, 7-20.
- Endicott, P. & Ho, S. Y. 2008. A Bayesian evaluation of human mitochondrial substitution rates. *Am J Hum Genet*, 82, 895-902.
- Endicott, P., Ho, S. Y., Metspalu, M. & Stringer, C. 2009. Evaluating the mitochondrial timescale of human evolution. *Trends Ecol Evol*, 24, 515-21.
- Excoffier, L. & Smouse, P. E. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics*, 136, 343-59.
- FamilyTreeDNA. 2006. *Armenian DNA Project - News* [Online]. Available: <http://www.familytreedna.com/public/armeniadnaproject/default.aspx?section=news> [Accessed 31/05/2013 2013].
- Fernandes, V., Alshamali, F., Alves, M., Costa, M. D., Pereira, J. B., Silva, N. M., Cherni, L., Harich, N., Cerny, V., Soares, P., Richards, M. B. & Pereira, L. 2012. The Arabian Cradle: Mitochondrial Relicts of the First Steps along the Southern Route out of Africa. *Am J Hum Genet*, 90, 347-55.
- Fernandez-Silva, P., Enriquez, J. A. & Montoya, J. 2003. Replication and transcription of mammalian mitochondrial DNA. *Exp Physiol*, 88, 41-56.

- Filosto, M., Mancuso, M., Vives-Bauza, C., Vila, M. R., Shanske, S., Hirano, M., Andreu, A. L. & DiMauro, S. 2003. Lack of paternal inheritance of muscle mitochondrial DNA in sporadic mitochondrial myopathies. *Ann Neurol*, 54, 524-6.
- Fix, A. 2011. Origin of genetic diversity among Malaysian Orang Asli: An alternative to the demic diffusion model. In: Enfield, N. J. (ed.) *Dynamics of human diversity: the case of mainland Southeast Asia*. Canberra, Australia: Pacific Linguistics.
- Fornarino, S., Pala, M., Battaglia, V., Maranta, R., Achilli, A., Modiano, G., Torroni, A., Semino, O. & Santachiara-Benerecetti, S. 2009. Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol Biol*, 9, 154.
- Forster, P. 2004. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Phil Trans R Soc Lond B*, 359, 255-64; discussion 264.
- Forster, P., Harding, R., Torroni, A. & Bandelt, H. J. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet*, 59, 935-45.
- Francalacci, P., Morelli, L., Angius, A., Berutti, R., Reinier, F., Atzeni, R., Pili, R., Busonero, F., Maschio, A., Zara, I., Sanna, D., Useli, A., Urru, M. F., Marcelli, M., Cusano, R., Oppo, M., Zoledziewska, M., Pitzalis, M., Deidda, F., Porcu, E., Poddie, F., Kang, H. M., Lyons, R., Tarrier, B., Gresham, J. B., Li, B., Tofanelli, S., Alonso, S., Dei, M., Lai, S., Mulas, A., Whalen, M. B., Uzzau, S., Jones, C., Schlessinger, D., Abecasis, G. R., Sanna, S., Sidore, C. & Cucca, F. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science*, 341, 565-9.
- Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G., Zhadanov, S., Schurr, T. G. & Merriwether, D. A. 2007. Melanesian mtDNA complexity. *PLoS One*, 2, e248.
- Fucharoen, G., Fucharoen, S. & Horai, S. 2001. Mitochondrial DNA polymorphisms in Thailand. *J Hum Genet*, 46, 115-25.
- Fullagar, R. L. K., Price, D. M. & Hea, L. M. 1996. Early human occupation of northern Australia: archaeology and thermoluminescence dating of Jinmium rock-shelter, Northern Territory. *Antiquity*, 70, 751-773.
- Gajra, B., Candlish, J. K., Heng, C. K., Mak, J. W. & Saha, N. 1997. Genotype associations among seven apolipoprotein B polymorphisms in a population of Orang Asli of western Malaysia. *Hum Bio*, 69, 629-640.
- Gajra, B., Candlish, J. K., Saha, N., Mak, J. W. & Tay, J. S. H. 1994. Effect of apolipoprotein E variants on plasma lipids and apolipoproteins in the Orang Asli ('aborigines') of Malaysia. *Hum Hered*, 44, 209-213.
- Gamble, C., Davies, W., Pettitt, P. & Richards, M. 2004. Climate change and evolving human diversity in Europe during the last glacial. *Phil Trans R Soc Lond B*, 359, 243-254.
- Goodwin, W. & Ovchinnikov, I. 2006. Ancient DNA and the Neanderthals. In: Bandelt, H.-J., Macaulay, V. & Richards, M. (eds.) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer Berlin Heidelberg.
- Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science*, 323, 479-483.
- Gray, R. D. & Jordan, F. M. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405, 1052-1055.
- Green, R. E., Malaspina, A. S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prufer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajkovic, D., Kucan, Z., Gusic,



- I., Wikstrom, M., Laakkonen, L., Kelso, J., Slatkin, M. & Paabo, S. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134, 416-26.
- Greenberg, J. H. 1971. The Indo-Pacific hypothesis. In: Sebeok, T. A. (ed.) *Current trends in linguistics*. Paris: Mouton, The Hague.
- Greenhill, S. J. & Gray, R. D. 2005. Testing Population Dispersal Hypotheses: Pacific Settlement, Phylogenetic Trees, and Austronesian Languages. In: Mace, R., Holden, C. & Shennan, S. (eds.) *The Evolution of Cultural Diversity: A Phylogenetic Approach*. California: Left Coast Press.
- Grine, F. E. & Henshilwood, C. S. 2002. Additional human remains from Blombos Cave, South Africa: (1999-2000 excavations). *J Hum Evol*, 42, 293-302.
- Guillot, E. G., Tumonggor, M. K., Lansing, J. S., Sudoyo, H. & Cox, M. P. 2013. Climate change influenced female population sizes through time across the Indonesian archipelago. *Hum Biol*, 85, 135-52.
- Gunnarsdóttir, E. D., Li, M., Bauchet, M., Finstermeier, K. & Stoneking, M. 2011a. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res*, 21, 1-11.
- Gunnarsdóttir, E. D., Nandineni, M. R., Li, M., Myles, S., Gil, D., Pakendorf, B. & Stoneking, M. 2011b. Larger mitochondrial DNA than Y-chromosome differences between matrilineal and patrilineal groups from Sumatra. *Nat Commun*, 2, 228.
- Hage, P. & Marck, J. 2003. Matrilineality and the Melanesian Origin of Polynesian Y Chromosomes. *Current Anthropology*, 44, S121-S127.
- Hagelberg, E., Goldman, N., Lio, P., Whelan, S., Schiefenhovel, W., Clegg, J. B. & Bowden, D. K. 1999. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc Biol Sci*, 266, 485-92.
- Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, A. S., Soodyall, H. & Zegura, S. L. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol*, 18, 1189-1203.
- Hanebuth, T., Stattegger, K. & Grootes, P. M. 2000. Rapid flooding of the sunda shelf: A late-glacial sea-level record. *Science*, 288, 1033-5.
- Hanihara, T. & Ishida, H. 2005. Metric dental variation of major human populations. *Am J Phys Anthropol*, 128, 287-298.
- Hartmann, A., Thieme, M., Nanduri, L. K., Stempf, T., Moehle, C., Kivisild, T. & Oefner, P. J. 2009. Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum Mutat*, 30, 115-122.
- Hasegawa, M., Kishino, H. & Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22, 160-174.
- Haslam, M., Clarkson, C., Petraglia, M., Korisettar, R., Jones, S., Shipton, C., Ditchfield, P. & Ambrose, S. H. 2010. The 74 ka Toba super-eruption and southern Indian hominins: archaeology, lithic technology and environments at Jwalapuram Locality 3. *Journal of Archaeological Science*, 37, 3370-3384.
- Hastings, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 97-109.
- Hatin, W. I., Nur-Shafawati, A. R., Zahri, M. K., Xu, S., Jin, L., Tan, S. G., Rizman-Idid, M. & Zilfalil, B. A. 2011. Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. *PLoS One*, 6, e18312.
- He, J. D., Peng, M. S., Quang, H. H., Dang, K. P., Trieu, A. V., Wu, S. F., Jin, J. Q., Murphy, R. W., Yao, Y. G. & Zhang, Y. P. 2012. Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. *PLoS One*, 7, e36437.

- Henn, B. M., Gignoux, C. R., Feldman, M. W. & Mountain, J. L. 2009. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol Biol Evol*, 26, 217-30.
- Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M., Anderson, C., Ghosh, S. S., Olefsky, J. M., Beal, M. F., Davis, R. E. & Howell, N. 2002a. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet*, 70, 1152-71.
- Herrnstadt, C., Preston, G., Andrews, R., Chinnery, P., Lightowlers, R. N., Turnbull, D. M., Kubacka, I. & Howell, N. 2002b. A high frequency of mtDNA polymorphisms in HeLa cell sublines. *Mutat Res*, 501, 19-28.
- Heyer, E., Georges, M., Pachner, M. & Endicott, P. 2013. Genetic diversity of four Filipino negrito populations from Luzon: comparison of male and female effective population sizes and differential integration of immigrants into Aeta and Agta communities. *Hum Biol*, 85, 189-208.
- Higham, C. F. W. 2004. Mainland Southeast Asia from the Neolithic to the Iron Age. In: Glover, I. & Bellwood, P. (eds.) *Southeast Asia: From Prehistory to History*. London: Routledge Curson.
- Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P. B., Vizuete-Forster, M., Forster, P., Bulbeck, D., Oppenheimer, S. & Richards, M. 2007. A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet*, 80, 29-43.
- Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., Clarke, D., Raja, J. M., Ismail, P., Bulbeck, D., Oppenheimer, S. & Richards, M. 2006. Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol*, 23, 2480-2491.
- Hixson, J. E. & Brown, W. M. 1986. A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. *Mol Biol Evol*, 3, 1-18.
- Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y., Fucharoen, G., Harihara, S., Park, K. S., Omoto, K. & Pan, I. H. 1996. mtDNA Polymorphism in East Asian Populations, with Special Reference to the Peopling of Japan. *Am J Hum Genet*, 59, 579-90.
- Howell, N., Smejkal, C. B., Mackey, D. A., Chinnery, P. F., Turnbull, D. M. & Herrnstadt, C. 2003. The Pedigree Rate of Sequence Divergence in the Human Mitochondrial Genome: There Is a Difference Between Phylogenetic and Pedigree Rates. *Am J Hum Genet*, 72, 659-70.
- Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., Shen, P., Oefner, P., Renfrew, C., Villems, R. & Forster, P. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Nat Acad Sci USA*, 104, 8726-30.
- Hung, H. C. 2005. Neolithic interaction between Taiwan and northern Luzon: the pottery and jade evidences from the Cagayan Valley. *Journal of Austronesian Studies*, 1, 109-134.
- Hung, H. C. 2008. *Migration and cultural interaction in southern coastal China, Taiwan and the Northern Philippines, 3000 BC to AD 1*. PhD, Australian National University.
- Hunt, C. O., Gilbertson, D. D. & Rushworth, G. 2007. Modern humans in Sarawak, Malaysian Borneo, during Oxygen Isotope Stage 3: palaeoenvironmental evidence from the Great Cave of Niah. *Journal of Archaeological Science*, 34, 1953-1969.
- Huntley, D. J., Godfrey-Smith, D. I. & Thewalt, M. L. W. 1985. Optical dating of sediments. *Nature*, 313, 105-107.

- Ingman, M. & Gyllensten, U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res*, 13, 1600-1606.
- Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408, 708-13.
- Innan, H. & Nordborg, M. 2002. Recombination or mutational hot spots in human mtDNA? *Mol Biol Evol*, 19, 1122-7.
- Ipoi, D. 1993. *Archaeological excavations at Gua Sireh (Serian) and Lubang Angin (Gunung Mulu National Park), Sarawak, Malaysia*, Sarawak, Sarawak Museum.
- Irwin, G. 1992. *The Prehistoric Exploration and Colonisation of the Pacific*, Cambridge, Cambridge University Press.
- Ivanoff, J. 2005. Sea Gypsies of Myanmar. *National Geographic*.
- Jiao, T. 2007. *The Neolithic of southeast China*, Youngstown, New York, Cambria.
- Jinam, T. A., Hong, L.-C., Phipps, M. E., Stoneking, M., Ameen, M., Edo, J., Consortium, H. P.-A. S. & Saitou, N. 2012. Evolutionary History of Continental Southeast Asians: “Early Train” Hypothesis Based on Genetic Analysis of Mitochondrial and Autosomal DNA Data. *Mol Biol Evol*, 29, 3513-3527.
- Jinam, T. A., Phipps, M. E. & Saitou, N. 2013. Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data. *Hum Biol*, 85, 173-88.
- Jobling, M., Hurles, M. & Tyler-Smith, C. 2003. *Human Evolutionary Genetics: Origins, Peoples and Disease*, New York, Garland Science.
- Karafet, T. M., Hallmark, B., Cox, M. P., Sudoyo, H., Downey, S., Lansing, J. S. & Hammer, M. F. 2010. Major east-west division underlies Y chromosome stratification across Indonesia. *Mol Biol Evol*, 27, 1833-44.
- Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. & Hammer, M. F. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18, 830-838.
- Kayser, M., Choi, Y., van Oven, M., Mona, S., Brauer, S., Trent, R. J., Suarkia, D., Schiefenhovel, W. & Stoneking, M. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol*, 25, 1362-74.
- Khrapko, K. 2008. Two ways to make an mtDNA bottleneck. *Nat Genet*, 40, 134-5.
- Kirch, P. V. 1997. *The Lapita Peoples: Ancestors of the Oceanic World*, Oxford, Blackwell.
- Kirch, P. V. 2000. *On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact*, California, University of California Press.
- Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E. & Villems, R. 2004. Ethiopian mitochondrial DNA heritage, tracking gene flow across and around the gate of tears. *Am J Hum Genet*, 75, 752-770.
- Kivisild, T., Roosti, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H. V., Stepanov, V., Golge, M., Usanga, E., Papiha, S. S., Cinnioglu, C., King, R., Cavalli Sforza, L., Underhill, P. A. & Villems, R. 2003. The genetic heritage of earliest settlers persist in both the Indian tribal and caste populations. *Am J Hum Genet*, 72, 313-332.
- Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P. A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knijff, P., Feldman, M., Cavalli-Sforza, L. L. & Oefner, P. J. 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics*, 172, 373-387.

- Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H. J. & Villems, R. 2002. The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol*, 19, 1737-1751.
- Kivisild, T., Villems, R., Jorde, L. B., Bamshad, M., Kumar, S., Hedrick, P., Dowling, T., Stoneking, M., Parsons, T. J., Irwin, J. A., Awadalla, P., Eyre-Walker, A. & Smith, J. M. 2000. Questioning evidence for recombination in human mitochondrial DNA. *Science*, 288, 1931.
- Klein, R. G. 1999. *The Human Career: Human Biological and Cultural Origins*, Chicago, University of Chicago Press.
- Kolman, C. J., Sambuughin, N. & Bermingham, E. 1996. Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics*, 142, 1321-34.
- Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Wong, W. S., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., Thorsteinsdottir, U. & Stefansson, K. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488, 471-5.
- Kong, Q. P., Bandelt, H. J., Sun, C., Yao, Y. G., Salas, A., Achilli, A., Wang, C. Y., Zhong, L., Zhu, C. L., Wu, S. F., Torroni, A. & Zhang, Y. P. 2006. Updating the East Asian mtDNA phylogeny, a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet*, 15, 2076-2086.
- Kong, Q. P., Sun, C., Wang, H. W., Zhao, M., Wang, W. Z., Zhong, L., Hao, X. D., Pan, H., Wang, S. Y., Cheng, Y. T., Zhu, C. L., Wu, S. F., Liu, L. N., Jin, J. Q., Yao, Y. G. & Zhang, Y. P. 2011. Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the peopling of the region. *Mol Biol Evol*, 28, 513-22.
- Kong, Q. P., Yao, Y. G., Liu, M., Shen, S. P., Chen, C., Zhu, C. L., Palanichamy, M. G. & Zhang, Y. P. 2003a. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Hum Genet*, 113, 391-405.
- Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L. & Zhang, Y. P. 2003b. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet*, 73, 671-676.
- Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Paabo, S. 1999. DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. *Proc Nat Acad Sci USA*, 96, 5581-5.
- Kumar, S. 2005. Molecular clock: four decades of evolution. *Nat Rev Genet*, 6, 654-662.
- Kumar, S., Padmanabham, P. B., Ravuri, R. R., Uttaravalli, K., Koneru, P., Mukherjee, P. A., Das, B., Kotal, M., Xaviour, D., Saheb, S. Y. & Rao, V. R. 2008. The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. *BMC Evol Biol*, 8, 230.
- Lahr, M. M. & Foley, R. 1994. Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews*, 3, 48-60.
- Lambeck, K. & Chappell, J. 2001. Sea level change through the last glacial cycle. *Science*, 292, 679-686.
- Larish, M. D. 1999. *The position of Moken and Moklen within the Austronesian language family*. Ph.D, University of Hawaii at Manoa.

- Li, K. C. 1983. *Report of archaeological investigations in the O-Luan-Pi Park at the southern tip of Taiwan*, Taipei, Department of Anthropology, National Taiwan University.
- Li, S. B. 2006. *GenBank Direct Submission* [Online]. Shaanxi, China: The State Key Laboratory of Forensic Sciences, College of Medicine, Xi'an Jiaotong University. Available: <http://www.ncbi.nlm.nih.gov/nuccore/DQ519035> [Accessed 01/02/2012 2012].
- Lin, M., Chu, C.-C., Broadberry, R. E., Yu, L.-C., Loo, J. H. & Trejaut, J. 2005. Genetic diversity of Taiwan's indigenous peoples: possible relationship with insular Southeast Asia. In: Sagart, L., Blench, R. & Sanchez-Mazas, A. (eds.) *The peopling of East Asia*. Abingdon (VA): RoutledgeCurzon.
- Loo, J.-H., Trejaut, J., Yen, J.-C., Chen, Z.-S., Lee, C.-L. & Lin, M. 2011. Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. *BMC Genet*, 12, 21.
- Louys, J., Curnoe, D. & Tong, H. 2007. Characteristics of Pleistocene megafauna extinctions in Southeast Asia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 243, 152-173.
- Luis, J. R., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioglu, C., Roseman, C., Underhill, P. A., Cavalli-Sforza, L. L. & Herrera, R. J. 2004. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am J Hum Genet*, 74, 532-44.
- Lum, J. K., Cann, R. L., Martinson, J. J. & Jorde, L. B. 1998. Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *Am J Hum Genet*, 63, 613-24.
- Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet*, 2, 13.
- Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N. K., Raja, J. M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H., Oppenheimer, S., Torroni, A. & Richards, M. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308, 1034-1036.
- Macaulay, V. & Richards, M. 2001. *Mitochondrial Genome Sequences and Their Phylogeographic Interpretation*. eLS. John Wiley & Sons, Ltd.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B. & Torroni, A. 1999a. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet*, 64, 232-49.
- Macaulay, V., Richards, M. & Sykes, B. 1999b. Mitochondrial DNA recombination-no need to panic. *Proc Biol Sci*, 266, 2037-9; discussion 2041-2.
- Margulis, L. 1981. *Symbiosis in cell evolution: life and its environment on the early Earth*, New York, W. H. Freeman & Co.
- Matsumura, H. & Hudson, M. J. 2004. Dental Perspectives on the Population History of Southeast Asia. *Am J Phys Anthropol*, 127, 182-209.
- McAllister, P., Nagle, N. & Mitchell, R. J. 2013. The Australian Barrineans and their relationship to Southeast Asian negritos: an investigation using mitochondrial genomics. *Hum Biol*, 85, 485-94.
- McBride, H. M., Neuspiel, M. & Wasiak, S. 2006. Mitochondria: more than just a powerhouse. *Curr Biol*, 16, R551-R560.

- McDougall, I., Brown, F. H. & Fleagle, J. G. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433, 733-6.
- McVean, G. A., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M. & Marth, G. T. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56-65.
- Meacham, W. 1978. *Sham Wan, Lamma Island: An Archaeological Site Study*, Hong Kong, Hong Kong Archaeological Society.
- Mellars, P. 2006. Going East: New Genetic and Archaeological Perspectives on the Modern Human Colonization of Eurasia. *Science*, 313, 796-800.
- Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. 2013. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc Nat Acad Sci USA*, 110, 10699-704.
- Mendez, F. L., Krahn, T., Schrack, B., Krahn, A. M., Veeramah, K. R., Woerner, A. E., Fomine, F. L., Bradman, N., Thomas, M. G., Karafet, T. M. & Hammer, M. F. 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am J Hum Genet*, 92, 454-9.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21, 1087-1092.
- Metspalu, M., Kivisild, T., Bandelt, H.-J., Richards, M. & Villems, R. 2006. The Pioneer Settlement of Modern Humans in Asia. In: Bandelt, H.-J., Macaulay, V. & Richards, M. (eds.) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer Berlin Heidelberg.
- Migliano, A. B., Vinicius, L. & Lahr, M. M. 2007. Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences*, 104, 20216-20219.
- Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M. D., Sukernik, R. I., Olckers, A. & Wallace, D. C. 2003. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences*, 100, 171-176.
- Mokhtar, S. 2006. Bukit Bunuh, Lenggong, Malaysia: new evidence of Late Pleistocene culture in Malaysia and Southeast Asia. In: Bacus, E. A., Glover, I. & Pigott, V. (eds.) *Uncovering Southeast Asia's Past*. Singapore: National University of Singapore Press.
- Mokhtar, S. & Tjia, H. D. 1994. Gua Gunung Runtuh: The Cave. In: Zuraina, M. (ed.) *The excavation of Gua Gunung Runtuh and the discovery of Perak Man in Malaysia*. Kuala Lumpur: Department of Museums and Antiquity Malaysia.
- Molengraaff, G. A. F. 1921. Modern Deep-Sea Research in the East Indian Archipelago. *The Geographical Journal*, 57, 95-118.
- Mormina, M. 2007. *mtDNA Perspectives on Human Dispersals in Mainland Southeast Asia*. PhD, University of Leeds.
- Morwood, M. J., Soejono, R. P., Roberts, R. G., Sutikna, T., Turney, C. S., Westaway, K. E., Rink, W. J., Zhao, J. X., van den Bergh, G. D., Due, R. A., Hobbs, D. R., Moore, M. W., Bird, M. I. & Fifield, L. K. 2004. Archaeology and age of a new hominin from Flores in eastern Indonesia. *Nature*, 431, 1087-91.
- Mountain, J. L., Hebert, J. M., Bhattacharyya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M. & Cavalli-Sforza, L. L. 1995. Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet*, 56, 979-92.

- Mudar, K. & Anderson, D. D. 2007. New evidence for Southeast Asian Pleistocene foraging economies: faunal remains from the early levels of Lang Rongrien rockshelter, Krabi, Thailand. *Asian Perspectives*, 46, 298-334.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. & Kent, J. 2000. Biodiversity hotspots for conservation priorities. *Nature*, 403, 853-8.
- Nishimaki, Y., Sato, K., Fang, L., Ma, M., Hasekura, H. & Boettcher, B. 1999. Sequence polymorphism in the mtDNA HV1 region in Japanese and Chinese. *Leg Med (Tokyo)*, 1, 238-49.
- Nohira, C., Maruyama, S. & Minaguchi, K. 2010. Phylogenetic classification of Japanese mtDNA assisted by complete mitochondrial DNA sequences. *Int J Legal Med*, 124, 7-12.
- Nur Haslindawaty, A. R., Panneerchelvam, S., Edinur, H. A., Norazmi, M. N. & Zafarina, Z. 2010. Sequence polymorphisms of mtDNA HV1, HV2, and HV3 regions in the Malay population of Peninsular Malaysia. *Int J Legal Med*, 124, 415-26.
- O'Connell, J. F. & Allen, J. 2004. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): a review of recent research. *Journal of Archaeological Science*, 31, 835-853.
- O'Connor, S., Spriggs, M. & Veth, P. 2007. *The Archaeology of the Aru Islands, Eastern Indonesia*, ANU E Press.
- Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N., Scozzari, R., Cruciani, F., Behar, D. M., Dugoujon, J. M., Coudray, C., Santachiara-Benerecetti, A. S., Semino, O., Bandelt, H. J. & Torroni, A. 2006. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science*, 314, 1767-1770.
- Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T. & Stoneking, M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet*, 29, 20-21.
- Oppenheimer, S. 1998. *Eden in the East*, London, Phoenix.
- Oppenheimer, S. 1999. *Eden in the East: the drowned continent of Southeast Asia*, London, Phoenix.
- Oppenheimer, S. 2003. *Out of Eden*, London, Constable and Robinson Ltd. .
- Oppenheimer, S. 2011. MtDNA variation and southward Holocene human dispersals within Mainland Southeast Asia. In: Enfield, N. J. (ed.) *The Dynamics of Human Diversity: The Case of Mainland Southeast Asia*. Canberra: Pacific Linguistics.
- Ovchinnikov, I. V., Gotherstrom, A., Romanova, G. P., Kharitonov, V. M., Liden, K. & Goodwin, W. 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, 404, 490-3.
- Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H. J., Kong, Q. P., Khan, F., Wang, C. Y., Chaudhuri, T. K., Palla, V. & Zhang, Y. P. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing, implications for the peopling of South Asia. *Am J Hum Genet*, 75, 966-978.
- Pawley, A. 1999. Chasing rainbows: implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In: Zeitoun, E. & Li, P. J. (eds.) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Taipei: Institute of Linguistics, Academia Sinica.
- Pawley, A. 2002. The Austronesian dispersal: languages, technologies and people. In: Bellwood, P. & Renfrew, C. (eds.) *Examining the Farming/Languages Dispersal Hypothesis*. Cambridge: McDonald Institute for Archaeological Research.

- Pelejero, C., Kienast, M., Wang, L. & Grimalt, J. O. 1999. The flooding of Sundaland during the last deglaciation: imprints in hemipelagic sediments from the southern South China Sea. *Earth and Planetary Science Letters*, 171, 661-671.
- Peng, M. S., He, J. D., Liu, H. X. & Zhang, Y. P. 2011b. Tracing the legacy of the early Hainan Islanders--a perspective from mitochondrial DNA. *BMC Evol Biol*, 11, 46.
- Peng, M. S., Palanichamy, M. G., Yao, Y. G., Mitra, B., Cheng, Y. T., Zhao, M., Liu, J., Wang, H. W., Pan, H., Wang, W. Z., Zhang, A. M., Zhang, W., Wang, D., Zou, Y., Yang, Y., Chaudhuri, T. K., Kong, Q. P. & Zhang, Y. P. 2011a. Inland post-glacial dispersal in East Asia revealed by mitochondrial haplogroup M9a'b. *BMC Biol*, 9, 2.
- Peng, M. S., Quang, H. H., Dang, K. P., Trieu, A. V., Wang, H. W., Yao, Y. G., Kong, Q. P. & Zhang, Y. P. 2010. Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. *Mol Biol Evol*, 27, 2417-30.
- Penny, D., Steel, M., Waddell, P. J. & Hendy, M. D. 1995. Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. *Mol Biol Evol*, 12, 863-82.
- Pereira, L., Soares, P., Radivojac, P., Li, B. & Samuels, D. C. 2011. Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. *Am J Hum Genet*, 88, 433-9.
- Pfeiffer, H., Steighner, R., Fisher, R., Mørnstad, H., Yoon, C. L. & Holland, M. M. 1998. Mitochondrial DNA extraction and typing from isolated dentin-experimental evaluation in a Korean population. *Int J Legal Med*, 111, 309-313.
- Pierson, M. J., Martinez-Arias, R., Holland, B. R., Gemmell, N. J., Hurles, M. E. & Penny, D. 2006. Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Mol Biol Evol*, 23, 1966-75.
- Pietrusewsky, M. 1997. The people of Ban Chiang: an early Bronze Age site in Northeast Thailand. *Bulletin of the Indo-Pacific Prehistory Association*, 119-147.
- Piganeau, G. & Eyre-Walker, A. 2004. A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. *Heredity (Edinb)*, 92, 282-8.
- Ponce de Leon, M. S. & Zollikofer, C. P. 2001. Neanderthal cranial ontogeny and its implications for late hominid diversity. *Nature*, 412, 534-8.
- Pradutkanchana, S., Ishida, T. & Kimura, R. 2010. Mitochondrial diversity of the sea nomads of Thailand. Songkhla, Thailand.
- Prasad, B. V., Ricker, C. E., Watkins, W. S., Dixon, M. E., Rao, B. B., Naidu, J. M., Jorde, L. B. & Bamshad, M. 2001. Mitochondrial DNA variation in Nicobarese Islanders. *Hum Bio*, 73, 715-25.
- Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A. S., Coppa, A., Ayub, Q., Mohyuddin, A., Tyler-Smith, C., Qasim Mehdi, S., Torroni, A. & McElreavey, K. 2004. Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor. *Am J Hum Genet*, 74, 827-845.
- Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. 1999. Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet*, 23, 437-41.
- Quirino, K. 2010. The Magnificent Journey To Aotearoa. *Karl Quirino's Blog: Keeping The World Hooked On To New Zealand* [Online]. Available from: <http://karlquirino.wordpress.com/2010/04/20/the-magnificent-journey-to-aotearoa/extent-of-austronesian-migrations/> [15/06/2013].



- Rajkumar, R., Banerjee, J., Gunturi, H. B., Trivedi, R. & Kashyap, V. K. 2005. Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. *BMC Evol Biol*, 5, 26.
- Rambo, A. T. 1988. Why are the Semang? Ecology and ethnogenesis of aboriginal groups in Peninsular Malaysia. In: Rambo, A. T., Gillogly, K. & Hutterer, K. L. (eds.) *Ethnic Diversity and the Control of Natural Resources in Southeast Asia*. Ann Arbor (MI): Center for South and Southeast Asian Studies, University of Michigan.
- Ramsey, C. B. 2008. Radiocarbon Dating: Revolutions in Understanding. *Archaeometry*, 50, 249-275.
- Rani, D. S., Dhandapany, P. S., Nallari, P., Govindaraj, P., Singh, L. & Thangaraj, K. 2010. Mitochondrial DNA haplogroup 'R' is associated with Noonan syndrome of south India. *Mitochondrion*, 10, 166-73.
- Razafindrazaka, H., Ricaut, F. X., Cox, M. P., Mormina, M., Dugoujon, J. M., Randriamarolaza, L. P., Guitard, E., Tonasso, L., Ludes, B. & Crubezy, E. 2010. Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *Eur J Hum Genet*, 18, 575-81.
- Redd, A. J., Takezaki, N., Sherry, S. T., McGarvey, S. T., Sofro, A. S. M. & Stoneking, M. 1995. Evolutionary history of the COII/tRNA<sup>Lys</sup> intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol*, 12, 604-615.
- Reid, L. A. 1994. Unravelling the linguistic histories of Philippine Negritos. In: Dutton, T. & Tyron, D. T. (eds.) *Language, contact and change in the Austronesian world*. Berlin: Moulton de Gruyter.
- Richards, M., Bandelt, H., Kivisild, T. & Oppenheimer, S. 2006. A model for the dispersal of modern humans out of Africa. In: Bandelt, H., Macaulay, V. & Richards, M. (eds.) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Berlin: Springer-Verlag.
- Richards, M., Macaulay, V. & Bandelt, H. J. 2002. Analyzing genetic data in a model-based framework: inferences about European prehistory. In: Renfrew, C. & Bellwood, P. S. (eds.) *Examining the Farming/Language Dispersal Hypothesis*. Cambridge, UK: McDonald Institute for Archaeological Research.
- Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L. & Galas, D. J. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328, 636-9.
- Roberts, R. G., Morwood, M. J. & Westaway, K. E. 2005. Illuminating Southeast Asian Prehistory: New Archaeological and Paleoanthropological Frontiers for Luminescence Dating. *Asian Perspectives*, 44, 293-319.
- Ross, M. D. 2008. *The integrity of the Austronesian language family: from Taiwan to Oceania*, Oxon, Routledge.
- Ruhlen, M. 1991. *A Guide to the World's Languages: Volume 1, Classification*, California, Stanford University Press.
- Sagart, L., Blench, R. & Sanchez-Mazas, A. 2005. *Introduction*, Oxon, RoutledgeCurzon.
- Saha, N., Mak, J. W., Tay, J. S. H., Liu, Y., Tan, J. A. M. A., Low, P. S. & Singh, M. 1995. Population Genetic Study among the Orang Asli (Semai Senoi) of Malaysia: Malayan Aborigines. *Hum Bio*, 67, 37.
- Saillard, J., Forster, P., Lynnerup, N., Bandelt, H. J. & Norby, S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet*, 67, 718-26.

- Salas, A., Richards, M., De la Fe, T., Lareu, M. V., Sobrino, B., Sanchez-Diz, P., Macaulay, V. & Carracedo, A. 2002. The making of the African mtDNA landscape. *Am J Hum Genet*, 71, 1082-111.
- Scarre, C. 2005. *The World Transformed: From Foragers and Farmers to States and Empires*, London, Thames and Hudson.
- Schebesta, P. & Blagden, C. O. 1926. The Jungle Tribes of the Malay Peninsula. *Bulletin of the School of Oriental and African Studies*, 4, 269-278.
- Scholes, C., Siddle, K., Ducourneau, A., Crivellaro, F., Jarve, M., Rootsi, S., Bellatti, M., Tabbada, K., Mormina, M., Reidla, M., Villems, R., Kivisild, T., Lahr, M. M. & Migliano, A. B. 2011. Genetic diversity and evidence for population admixture in Batak Negritos from Palawan. *Am J Phys Anthropol*, 146, 62-72.
- Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B. & Wallace, D. C. 1999. Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am J Phys Anthropol*, 108, 1-39.
- Schwartz, M. & Vissing, J. 2002. Paternal inheritance of mitochondrial DNA. *N Engl J Med*, 347, 576-80.
- Schwartz, M. & Vissing, J. 2004. No evidence for paternal inheritance of mtDNA in patients with sporadic mtDNA mutations. *J Neurol Sci*, 218, 99-101.
- Semino, O., Santachiara-Benerecetti, A. S., Falaschi, F., Cavalli-Sforza, L. L. & Underhill, P. A. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet*, 70, 265-8.
- Seo, Y., Stradmann-Bellinghausen, B., Rittner, C., Takahama, K. & Schneider, P. M. 1998. Sequence polymorphism of mitochondrial DNA control region in Japanese. *Forensic Sci Int*, 97, 155-64.
- Shang, H., Tong, H., Zhang, S., Chen, F. & Trinkaus, E. 2007. An early modern human from Tianyuan Cave, Zhoukoudian, China. *Proceedings of the National Academy of Sciences*, 104, 6573-6578.
- Sharma, G., Tamang, R., Chaudhary, R., Singh, V. K., Shah, A. M., Anugula, S., Rani, D. S., Reddy, A. G., Eaaswarkhanth, M., Chaubey, G., Singh, L. & Thangaraj, K. 2012. Genetic affinities of the central Indian tribal populations. *PLoS One*, 7, e32546.
- Shi, W., Ayub, Q., Vermeulen, M., Shao, R.-g., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y. & Tyler-Smith, C. 2010. A Worldwide Survey of Human Male Demographic History Based on Y-SNP and Y-STR Data from the HGDP-CEPH Populations. *Mol Biol Evol*, 27, 385-393.
- Sidwell, P. & Blench, R. 2011. The Austroasiatic Urheimat: the Southeastern Riverine Hypothesis. In: Enfield, N. J. (ed.) *Dynamics of human diversity: the case of mainland Southeast Asia*. Canberra Australia: Pacific Linguistics.
- Simonson, T. S., Xing, J., Barrett, R., Jerah, E., Loa, P., Zhang, Y., Watkins, W. S., Witherspoon, D. J., Huff, C. D., Woodward, S., Mowry, B. & Jorde, L. B. 2011. Ancestry of the Iban is predominantly Southeast Asian: genetic evidence from autosomal, mitochondrial, and Y chromosomes. *PLoS One*, 6, e16338.
- Skeat, W. W. & Blagden, C. O. 1906. *The Pagan Races of the Malay Peninsula*, London, Macmillan and Co., limited.
- Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., Costa, M. D., Musilova, E., Macaulay, V., Richards, M. B., Cerny, V. & Pereira, L. 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Mol Biol Evol*, 29, 915-27.
- Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., Salas, A., Oppenheimer, S., Macaulay, V. & Richards, M. B. 2009. Correcting for purifying

- selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*, 84, 740-59.
- Soares, P., Rito, T., Trejaut, J., Mormina, M., Hill, C., Tinkler-Hundal, E., Braid, M., Clarke, D. J., Loo, J. H., Thomson, N., Denham, T., Donohue, M., Macaulay, V., Lin, M., Oppenheimer, S. & Richards, M. B. 2011. Ancient voyaging and Polynesian origins. *Am J Hum Genet*, 88, 239-47.
- Soares, P., Trejaut, J. A., Loo, J. H., Hill, C., Mormina, M., Lee, C. L., Chen, Y. M., Hudjashov, G., Forster, P., Macaulay, V., Bulbeck, D., Oppenheimer, S., Lin, M. & Richards, M. B. 2008. Climate change and postglacial human dispersals in southeast Asia. *Mol Biol Evol*, 25, 1209-18.
- Solheim, W. G. I. 1980. Searching for the origins of the Orang Asli. *Federation Museums Journal*, 25, 61-75.
- Solheim, W. G. I. 2006. *Archaeology and Culture in Southeast Asia: Unraveling the Nusantara*, Quezon City, The Philippines, University of Philippines Press.
- Sopher, D. E. 1965. *The Sea Nomads: a Study Based on the Literature of the Maritime Boat People of Southeast Asia*, Singapore, Lim Bian Han Government Printer.
- Specht, J. 2005. *Revisiting the Bismarcks: some alternative views*, Canberra, Pacific Linguistics.
- Spriggs, M. 2007. The Neolithic and Austronesian Expansion within Island Southeast Asia and into the Pacific. In: Chiu, S. & Sand, C. (eds.) *From Southeast Asia to the Pacific. Archaeological Perspectives on the Austronesian Expansion and the Lapita Cultural Complex*. Taipei: Academia Sinica.
- Starikovskaya, E. B., Sukernik, R. I., Derbeneva, O. A., Volodko, N. V., Ruiz-Pesini, E., Torroni, A., Brown, M. D., Lott, M. T., Hosseini, S. H., Huoponen, K. & Wallace, D. C. 2005. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann Hum Genet*, 69, 67-89.
- Stock, J. T. 2013. The skeletal phenotype of "negritos" from the Andaman Islands and Philippines relative to global variation among hunter-gatherers. *Hum Biol*, 85, 67-94.
- Stoneking, M. & Delfin, F. 2010. The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol*, 20, R188-93.
- Stringer, C. 2000. Palaeoanthropology: Coasting out of Africa. *Nature*, 405, 24-27.
- Stringer, C. B. & Andrews, P. 1988. Genetic and fossil evidence for the origin of modern humans. *Science*, 239, 1263-8.
- Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., Huang, W., Shen, D., Lu, D., Luo, J., Chu, J., Tan, J., Shen, P., Davis, R., Cavalli-Sforza, L., Chakraborty, R., Xiong, M., Du, R., Oefner, P., Chen, Z. & Jin, L. 1999. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet*, 65, 1718-24.
- Summerer, M., Horst, J., Erhart, G., Weissensteiner, H., Schonherr, S., Pacher, D., Forer, L., Horst, D., Manhart, A., Horst, B., Sanguansermsri, T. & Kloss-Brandstatter, A. 2014. Large-scale mitochondrial DNA analysis in Southeast Asia reveals evolutionary effects of cultural isolation in the multi-ethnic population of Myanmar. *BMC Evol Biol*, 14, 17.
- Summerer, M., Horst, J. R., Erhart, G., Horst, D., Manhart, A., Horst, B., Sanguansermsri, T., Kronenberg, F. & Kloss-Brandstatter, A. 2012. Comparative mitochondrial DNA analyses of the Karen people and their distinct genetic position within the multi-ethnic population of Myanmar. 06/07/2012 ed. Bethesda MD: National Center for Biotechnology Information, U.S. National Library of Medicine.

- Summerhayes, G. R., Leavesley, M., Fairbairn, A., Mandui, H., Field, J., Ford, A. & Fullagar, R. 2010. Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. *Science*, 330, 78-81.
- Sun, C., Kong, Q. P., Palanichamy, M. G., Agrawal, S., Bandelt, H. J., Yao, Y. G., Khan, F., Zhu, C. L., Chaudhuri, T. K. & Zhang, Y. P. 2006. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol*, 23, 683-690.
- Swisher, C. C., 3rd, Curtis, G. H., Jacob, T., Getty, A. G., Suprijo, A. & Widiasmoro 1994. Age of the earliest known hominids in Java, Indonesia. *Science*, 263, 1118-21.
- Swisher, C. C., 3rd, Rink, W. J., Anton, S. C., Schwarcz, H. P., Curtis, G. H., Suprijo, A. & Widiasmoro 1996. Latest *Homo erectus* of Java: potential contemporaneity with *Homo sapiens* in southeast Asia. *Science*, 274, 1870-4.
- Sykes, B., Leiboff, A., Low-Beer, J., Tetzner, S. & Richards, M. 1995. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. *Am J Hum Genet*, 57, 1463-75.
- Tabbada, K. A., Trejaut, J., Loo, J. H., Chen, Y. M., Lin, M., Mirazon-Lahr, M., Kivisild, T. & De Ungria, M. C. 2010. Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? *Mol Biol Evol*, 27, 21-31.
- Tajima, A., Sun, C. S., Pan, I. H., Ishida, T., Saitou, N. & Horai, S. 2003. Mitochondrial DNA polymorphisms in nine aboriginal groups of Taiwan: implications for the population history of aboriginal Taiwanese. *Hum Genet*, 113, 24-33.
- Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D. G., Mulligan, C. J., Bravi, C. M., Rickards, O., Martinez-Labarga, C., Khusnutdinova, E. K., Fedorova, S. A., Golubenko, M. V., Stepanov, V. A., Gubina, M. A., Zhadanov, S. I., Ossipova, L. P., Damba, L., Voevoda, M. I., Dipierri, J. E., Villems, R. & Malhi, R. S. 2007. Beringian standstill and spread of Native American founders. *PLoS One*, 2, e829.
- Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T., Fuku, N., Guo, L. J., Hirose, R., Fujita, Y., Kurata, M., Shinoda, K., Umetsu, K., Yamada, Y., Oshida, Y., Sato, Y., Hattori, N., Mizuno, Y., Arai, Y., Hirose, N., Ohta, S., Ogawa, O., Tanaka, Y., Kawamori, R., Shamoto-Nagai, M., Maruyama, W., Shimokata, H., Suzuki, R. & Shimodaira, H. 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res*, 14, 1832-1850.
- Taylor, R. E. 1995. Radiocarbon dating: The continuing revolution. *Evolutionary Anthropology: Issues, News, and Reviews*, 4, 169-181.
- Taylor, R. W., McDonnell, M. T., Blakely, E. L., Chinnery, P. F., Taylor, G. A., Howell, N., Zeviani, M., Briem, E., Carrara, F. & Turnbull, D. M. 2003. Genotypes from patients indicate no paternal mitochondrial DNA contribution. *Ann Neurol*, 54, 521-4.
- Taylor, R. W. & Turnbull, D. M. 2005. Mitochondrial DNA mutations in human disease. *Nat Rev Genet*, 6, 389-402.
- Templeton, A. 1993. The "Eve" Hypotheses: A Genetic Critique and Reanalysis. *American Anthropologist*, 95, 51-72.
- Terberger, T. & Street, M. 2002. Hiatus or continuity? New results for the question of pleniglacial settlement in Central Europe. *Antiquity*, 76, 691-698.
- Terrell, J. E. 1986. *Prehistory in the Pacific Islands*, Cambridge, Cambridge University Press.
- Terrell, J. E. 2004. The 'sleeping giant' hypothesis and New Guinea's place in the prehistory of Greater Near Oceania. *World Archaeol*, 36, 601-609.
- Terrell, J. E., Kelly, K. M. & Rainbird, P. 2001. Foregone conclusions? In search of "Papuan" and "Austronesians". *Current Anthropology*, 42, 97-124.

- Terrell, J. E. & Welsch, R. L. 1997. Lapita and the temporal geography of prehistory. *Antiquity*, 71, 548-572.
- Thangaraj, K., Chaubey, G., Kivisild, T., Reddy, A. G., Singh, V. K., Rasalkar, A. A. & Singh, L. 2005. Reconstructing the origin of Andaman Islanders. *Science*, 308, 996.
- Thangaraj, K., Chaubey, G., Singh, V. K., Vanniarajan, A., Thanseem, I., Reddy, A. G. & Singh, L. 2005b. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. *BMC Genomics*, 7, 151.
- Thangaraj, K., Chaubey, G., Singh, V. K., Vanniarajan, A., Thanseem, I., Reddy, A. G. & Singh, L. 2006. In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India. *BMC Genomics*, 7, 151.
- Torrence, R. & Swadling, P. 2008. Social networks and the spread of Lapita. *Antiquity*, 82, 600-616.
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M. L. & Wallace, D. C. 1996. Classification of European mtDNAs from an analysis of three European populations. *Genetics*, 144, 1835-50.
- Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., Smith, D. G., Vullo, C. M. & Wallace, D. C. 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet*, 53, 563-90.
- Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., Lee, Z. Y. & Lin, M. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *PLoS Biol*, 3, e247.
- Tsang, C. H. 2005. *Recent discoveries at a Tapenkeng culture site in Taiwan: implications for the problem of Austronesian origins*, Oxon, RoutledgeCurzon.
- Tsang, C. H., Li, K. T. & Chu, C. Y. 2006. *Footprints of ancestors: archaeological discoveries in Tainan Science-Based Industrial Park*, Tainan, Tainan County Government.
- Turner, C. G. I. 1987. Late Pleistocene and Holocene population history of East Asia based on dental variation. *Am J Phys Anthropol*, 73, 305-321.
- Turney, C. S. M., Bird, M. I., Fifield, L. K., Roberts, R. G., Smith, M., Dortch, C. E., Grün, R., Lawson, E., Ayliffe, L. K., Miller, G. H., Dortch, J. & Cresswell, R. G. 2001. Early Human Occupation at Devil's Lair, Southwestern Australia 50,000 Years Ago. *Quaternary Research*, 55, 3-13.
- Underhill, P. A. & Kivisild, T. 2007a. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*. Palo Alto: Annual Reviews.
- Underhill, P. A. & Kivisild, T. 2007b. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annu Rev Genet*, 41, 539-564.
- Valentine, B., Kamenov, G. & Krigbaum, J. 2008. Reconstructing Neolithic groups in Sarawak, Malaysia through lead and strontium isotope analysis. *Journal of Archaeological Science*, 35, 1463-1473.
- van Heekeren, H. R. 1972. *The Stone Age of Indonesia*, Martinus Nijhoff, The Hague.
- van Oven, M. & Kayser, M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*, 30, E386-94.
- Vandamme, A.-M. 2009. *Basic concepts of molecular evolution*, New York, Cambridge University Press.
- Verstappen, H. T. 1997. The effect of climatic change on southeast Asian geomorphology. *Journal of Quaternary Science*, 12, 413-418.
- Voris, H. K. 2000. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. *Journal of Biogeography*, 27, 1153-1167.

- Wallace, A. R. 1881. *Island life; or, The phenomena and causes of insular faunas and floras, including a revision and attempted solution of the problem of geological climates*, New York, Harper & brothers.
- Wang, H. W., Mitra, B., Chaudhuri, T. K., Palanichamy, M. G., Kong, Q. P. & Zhang, Y. P. 2011. Mitochondrial DNA evidence supports northeast Indian origin of the aboriginal Andamanese in the Late Paleolithic. *J Genet Genomics*, 38, 117-22.
- Watson, E., Forster, P., Richards, M. & Bandelt, H. J. 1997. Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet*, 61, 691-704.
- Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., Jim, L., Su, B., Pitchappan, R., Shanmuglakshmi, S., Balakrishnan, K., Read, M., Pearson, N. M., Zerjal, T., Webster, M. T., Zholoshvili, I., Jamarjashvili, E., Gambarov, S., Nikbin, B., Dostiev, A., Aknazarov, O., Zalloua, P., Tsoy, I., Kitaev, M., Mirrakhimov, M., Chariev, A. & Bodmer, W. F. 2001. The Eurasian Heartland, A Continental Perspective on Y-chromosome Diversity. *Proc Nat Acad Sci USA*, 98, 10244-10249.
- Wen, B., Li, H., Gao, S., Mao, X., Gao, Y., Li, F., Zhang, F., He, Y., Dong, Y., Zhang, Y., Huang, W., Jin, J., Xiao, C., Lu, D., Chakraborty, R., Su, B., Deka, R. & Jin, L. 2005. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol*, 22, 725-34.
- Wheatley, P. 1961. *The Golden Khersonese: studies in the historical geography of the Malay Peninsula before A. D. 1500*, Kuala Lumpur, University of Malaya Press.
- White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G. & Howell, F. C. 2003. Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature*, 423, 742-7.
- White, W. G. 1922. *The Sea Gypsies of Malaya*, Philadelphia, JB Lippincott Co.
- Wickler, S. & Spriggs, M. 1988. Pleistocene occupation of the Solomon Islands, Melanesia. *Antiquity*, 62, 703-706.
- Wolpoff, M. & Thorne, A. 1991. The case against Eve. *New Scientist*, 130, 37-41.
- Wolpoff, M. H., Spuhler, J. N., Smith, F. H., Radovic, J., Pope, G., Frayer, D. W., Eckhardt, R. & Clark, G. 1988. Modern human origins. *Science*, 241, 772-4.
- Wong, L. P., Ong, R. T., Poh, W. T., Liu, X., Chen, P., Li, R., Lam, K. K., Pillai, N. E., Sim, K. S., Xu, H., Sim, N. L., Teo, S. M., Foo, J. N., Tan, L. W., Lim, Y., Koo, S. H., Gan, L. S., Cheng, C. Y., Wee, S., Yap, E. P., Ng, P. C., Lim, W. Y., Soong, R., Wenk, M. R., Aung, T., Wong, T. Y., Khor, C. C., Little, P., Chia, K. S. & Teo, Y. Y. 2013. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet*, 92, 52-66.
- Wurm, S. A. & Hattori, S. 1981. *Language atlas of the Pacific area [cartographic material]*, Canberra, Australian Academy of the Humanities in collaboration with the Japan Academy.
- Wurm, S. A. & McElhanan, K. 1975. *New Guinea area languages and language study, vol 1*, Canberra, Australian National University.
- Xu, X. & Arnason, U. 1996. A complete sequence of the mitochondrial genome of the western lowland gorilla. *Mol Biol Evol*, 13, 691-8.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13, 555-6.
- Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T. & Zhang, Y. P. 2002a. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet*, 70, 635-651.
- Yao, Y. G., Nie, L., Harpending, H., Fu, Y. X., Yuan, Z. G. & Zhang, Y. P. 2002b. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol*, 118, 63-76.

- Yao, Y. G. & Zhang, Y. P. 2002. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan province: new data and a reappraisal. *J Hum Genet*, 47, 311-318.
- Zainuddin, Z. & Goodwin, W. 2004. Mitochondrial DNA profiling of modern Malay and Orang Asli populations in Peninsular Malaysia. *International Congress Series*, 1261, 428-430.
- Zhang, X., Qi, X., Yang, Z., Serey, B., Sovannary, T., Bunnath, L., Seang Aun, H., Samnom, H., Zhang, H., Lin, Q., van Oven, M., Shi, H. & Su, B. 2013. Analysis of mitochondrial genome diversity identifies new and ancient maternal lineages in Cambodian aborigines. *Nat Commun*, 4, 2599.
- Zhao, M., Kong, Q. P., Wang, H. W., Peng, M. S., Xie, X. D., Wang, W. Z., Jiayang, Duan, J. G., Cai, M. C., Zhao, S. N., Cidanpingcuo, Tu, Y. Q., Wu, S. F., Yao, Y. G., Bandelt, H. J. & Zhang, Y. P. 2009. Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc Nat Acad Sci USA*, 106, 21230-5.
- Zheng, H.-X., Yan, S., Qin, Z.-D., Wang, Y., Tan, J.-Z., Li, H. & Jin, L. 2011. Major Population Expansion of East Asians Began before Neolithic Time: Evidence of mtDNA Genomes. *PLoS One*, 6, e25835.
- Zuraina, M. 1990. The Tampanian problem resolved: archaeological evidence of a late Pleistocene lithic workshop. *Modern Quaternary Research in Southeast Asia*, 11, 71-96.
- Zuraina, M., Ang, B. H. & Jeffri, I. 1998. Late Pleistocene-Holocene sites in Pahang: excavations of Gua Sagu and Gua Tenggek. *Malaysia Museums Journal*, 34, 65-115.
- Zuraina, M., Mokhtar, S., Chia, S. & Zolkurnian, H. 1994. Artifacts from Gua Gunung Runtuh excavation. In: Zuraina, M. (ed.) *The Excavation of Gua Gunung Runtuh and the Discovery of Perak Man in Malaysia*. Kuala Lumpur: Department of Museums and Antiquity Malaysia.
- Zuraina, M. & Tjia, H. D. 1988. Kota Tampan, Perak: The geological and archaeological evidence for a Late Pleistocene site. *Journal of the Malaysian Branch of the Royal Asiatic Society*, 61, 123-134.

## Appendices

### Appendix A

The three-letters codes used to show the locations or regions of the complete sequences in the phylogeography analysis.

Code	Location	No. of Complete Sequences
AFR	African	1
ALG	Algeria	1
AMC	American, Caucasian	1
AME	American, Native	6
AUS	Australia	11
BGL	Bangladesh	1
BIS	Bismarck Islands	23
BOU	Bougainville	5
BRA	Brazil	5
BRU	Brunei, North Borneo	2
CAM	Cambodia	4
CHI	China	303
MGL	China, Inner Mongolia	17
XIN	China, Xinjiang	14
COL	Colombia	1
COO	Cook Island	3
ESK	Eskimo	9
EUR	Europe	2
IBE	Europe, Iberian Peninsular	1
FRA	France	1
GEO	Georgia	1
IND	India	76
AND	India, Andaman Islands	5



HIM	India, Himalaya	1
NIC	India, Nicobars	5
INA	Indonesia	6
JAV	Indonesia, Java	20
LSI	Indonesia, Lesser Sunda Islands	18
MOL	Indonesia, Moluccas	5
SBO	Indonesia, South Borneo	19
SUL	Indonesia, Sulawesi	21
SUM	Indonesia, Sumatra	81
WNG	Indonesia, West New Guinea	2
ISR	Israel	2
ITA	Italy	3
JAP	Japan	670
KOR	Korea	2
KYR	Kyrgyzstan	1
MAD	Madagascar	1
ABM	Malaysia, Aboriginal Malay	1
JAK	Malaysia, Aboriginal Malay, Jakun	1
SEL	Malaysia, Aboriginal Malay, Seletar	21
SML	Malaysia, Aboriginal Malay, Semelai	11
TEM	Malaysia, Aboriginal Malay, Temuan	24
NBO	Malaysia, North Borneo	12
NEM	Malaysia, Northeast Peninsular Malays	73
NWM	Malaysia, Northwest Peninsular Malays	64
BID	Malaysia, Sarawak, Bidayuh	23
BAT	Malaysia, Semang, Batek	5
JAH	Malaysia, Semang, Jahai	26
KEN	Malaysia, Semang, Kensiu	7
KIN	Malaysia, Semang, Kintak	5
LAN	Malaysia, Semang, Lanoh	2
MEN	Malaysia, Semang, Mendriq	1

SMI	Malaysia, Senoi, Semai	1
TMI	Malaysia, Senoi, Temiar	7
SEM	Malaysia, Southeast Peninsular Malays	33
SWM	Malaysia, Southwest Peninsular Malays	18
MEX	Mexico	3
MIC	Micronesia	4
MYA	Myanmar	2
NEP	Nepal	5
PAK	Pakistan	10
PNG	Papua New Guinea	50
FIL	Philippines	123
FBT	Philippines, Batak	5
MAM	Philippines, Mamanwa	38
RUS	Russia	2
SBR	Russia, Siberian	28
SMO	Samoa	4
SRI	Sri Lanka	1
TAI	Taiwan	41
THA	Thailand	64
TMK	Thailand, Moken	24
TIB	Tibet	19
TON	Tonga	2
TUN	Tunisia	1
UZB	Uzbekistan	1
VAN	Vanuatu	14
VIE	Vietnam	80
Total		2206

## Appendix B

Variant positions for 91 *Orang Asli* HVS-I sequences analysed in Section 3.1. Sequences provided by K.C. Ang (unpublished data).

Samples	Haplogroup	HVS-I	Total
PMJakun4	B4c	16140 16189 16213 16217 16274 16335 16519	1
PMSemelai1	B5a	16140 16189 16261 16266A 16519	1
PMJakun2, PMJakun3, PNKanak2, PMSeletar2, PMSeletar3, PMSeletar4	E	16223 16291 16362 16390 16519	6
PMJakun5, PMSemelai3, PMSemelai5	F1a1a	16108 16129 16162 16172 16189 16304 16519	3
PMKuala2, PMKuala3	M	16189 16223 16278	2
PMKuala4	M	16214A 16223 16278	1
PMKuala1	M	16189 16223 16278 16354 16519	1
PMKuala5	M	16086 16223 16259 16278 16319 16399 16526	1
PNKanak1, PNKanak3, PNKanak4, PNKanak5	M	16093 16209 16223 16224 16263 16278 16319	4
NBatek1, NBatek2, NBatek3, NBatek4, NBatek5, SCheWong2, SCheWong4, SCheWong5	M21a	16129 16223 16256 16271 16362	8
SCheWong1, SCheWong3	M21a	16129 16140 16223 16256 16271 16362	2
PMTemuan5	M21a	16093 16129 16192iC 16223 16256 16362	1
NMendrik1, NMendrik2, NMendrik3, SJahHut1, SJahHut2, SJahHut3, SJahHut4, SJahHut5	M21a	16093 16129 16217 16223 16256 16271 16362	8
NMendrik4	M21a	16093 16100T 16129 16223 16256 16271 16284 16362	1
NMendrik5	M21b	16093 16129 16223 16256 16263 16381 16519	1
PMSemelai2	M7c3c	16223 16295 16362 16519	1
PMTemuan1, PMTemuan4	N21	16193 16291 16519	2
PMSemelai4	N21	16193 16291 16327 16519	1
PMTemuan2	N22	16168 16223 16249	1
PMSeletar1, PMSeletar5	N9a6	16223 16257A 16261 16292 16342 16519	2

PMTemuan3	R21	16295 16296 16304	1
NJahai1	R21	16086 16295 16296 16304	1
NLanoh1, NLanoh2, NLanoh6, SSemai1, SSemai2, SSemai3, SSemai5	R21	16168 16295 16296 16304	7
NLanoh3, Nlanoh4, NLanoh5, SSemai4	R21	16168 16295 16296 16304 16519	4
STemiar3	R21	16086 16168 16197A 16296 16304	1
SMahMeri2, SMahMeri3, SMahMeri4, SMahMeri5	R21	16086 16168 16192 16197A 16296 16304	4
SSemokBeri1, STemiar1, STemiar5	R21	16086 16168 16197A 16295 16296 16304	3
STemiar2, STemiar4	R21	16086 16168 16197A 16296 16304 16519	2
SMahMeri1	R21	16086 16168 16192 16197A 16296 16304 16519	1
SSemokBeri2, SSemokBeri3, SSemokBeri5	R21	16086 16168 16197A 16271 16295 16296 16304	3
SSemokBeri4	R21	16086 16168 16197A 16295 16296 16304 16519	1
NJahai2, NJahai3, NJahai4, NJahai5, NKensiu1, NKensiu2, NKensiu3, NKensiu4, NKensiu5, NKintak1, NKintak2, NKintak3, NKintak4, NKintak5	R21	16086 16168 16192G 16197A 16295 16296 16304	14
PMJakun1	R9b	16086 16170 16223 16288 16304 16309 16390	1

## Appendix C

List of 226 complete mtDNA genomes of *Orang Asli* and Peninsular Malays, and the world regional distributions of each complete sequence haplogroup.

Haplogroup	Semang	Senoi	Aboriginal Malay	Peninsular Malays	Total	Regional distribution in this study & review
A5b				1	1	China, Japan, SEM
B4a1a				7	7	ISEA
B4a1a1a				1	1	Oceania
B4a1c			1		1	Japan, China, North Asia
B4a1c+146				1	1	China, North Asia, MSEA
B4a1c4				1	1	China, North Asia, MSEA
B4b1a2				1	1	China, Japan, ISEA
B4c1b2a2				10	10	Taiwan, ISEA, Peninsular Malaysia
B4c2	2			2	4	MSEA
B4c2b		1		2	3	MSEA
B5a1a	1			9	10	SEA
B5a1b				1	1	China, Philippines, Peninsular Malaysia
B5a1d				2	2	SEA
B5b1c	1			2	3	South China
B6a1a			1	3	4	Peninsular Malaysia
C7a				1	1	MSEA, northern China
C7a1d				1	1	Peninsular Malaysia
D4a3b				1	1	Peninsular Malaysia, northern China
D5b3				1	1	MSEA
E1a1a				4	4	Taiwan, ISEA, Peninsular Malaysia
E1a1a1				1	1	SEA
E1a2				2	2	ISEA, Peninsular Malaysia, PNG
E1b+16261				3	3	ISEA, Peninsular Malaysia, PNG
E2a				2	2	ISEA, Peninsular Malaysia, PNG
E2a3				1	1	ISEA, Peninsular Malaysia
F1a1				2	2	South China, SEA
F1a1a		1		6	7	Peninsular Malaysia, ISEA, South China
F1a1a1		3	1	5	9	SEA
F1a1c				1	1	China, Japan, MSEA
F1a1d				1	1	SEA, Taiwan
F1a3a				1	1	SEA, Japan
F1a4a				1	1	SEA
F1f				4	4	SEA
F3a1				1	1	North China, Peninsular Malaysia

F3a2				1	1	North China, MSEA
F3b1a				1	1	SEA, Taiwan
F4b				1	1	China, India, Peninsular Malaysia
M*				1	1	
M12				1	1	MSEA
M12a1b1				1	1	MSEA, India
M12a2				1	1	MSEA
M12b1				2	2	SEA
M13'46'61+163 62				1	1	East Eurasia
M13b1			1	2	3	Peninsular Malaysia
M17a1a	2				2	MSEA
M17c				3	3	SEA
M20				1	1	MSEA
M20a				1	1	SEA
M20a1a				1	1	SEA
M20a1a1				3	3	SEA
M20a1a2				1	1	SEA
M21a1a				1	1	ISEA
M21a1b	7			1	8	MSEA
M21a1c			1	2	3	SEA
M21c				1	1	SEA
M21c1				1	1	SEA
M21c2				1	1	SEA
M21d1a				2	2	MSEA
M22a2				2	2	Peninsular Malaysia
M22b1				1	1	MSEA
M26a				1	1	SEA
M26a1				1	1	SEA
M26b				1	1	MSEA
M26b2				1	1	SEA
M2a'b				1	1	Peninsular Malaysia, Pakistan, Brazil
M30a1				1	1	India, Peninsular Malaysia
M32c				1	1	Peninsular Malaysia, Madagascar
M37e				1	1	India, Peninsular Malaysia
M4"67				1	1	SEA, Australia
M47				1	1	SEA
M50a				1	1	SEA
M50a1				1	1	SEA
M50a1a				1	1	SEA
M50b				2	2	MSEA
M51a2				1	1	SEA
M51b1a1				1	1	Vietnam
M51b2				1	1	MSEA
M51b2a				1	1	Vietnam
M5a				1	1	South Asia, Peninsular Malaysia
M71a1				1	1	MSEA
M71b				1	1	MSEA
M72a1a				1	1	MSEA
M72a1b				1	1	SEA
M73b1a				1	1	SEA

M74b1				1	1	SEA
M77				1	1	SEA
M7b1'2'4-8+16189				3	3	China, Japan, Peninsular Malaysia
M7b3a				1	1	East Eurasia
M7b7				1	1	North China, Peninsular Malaysia
M7c3c			1	6	7	ISEA, Taiwan, Micronesia
N10a1				1	1	SEA
N21			1	2	3	MSEA
N22a			1		1	Peninsular Malaysia
N22b				1	1	SEA
N8a				1	1	MSEA
N9a				1	1	China, Japan
N9a6				1	1	SEA
N9a6a	1		2		3	SEA
P1d1				2	2	PNG, Peninsular Malaysia
Q1+@16223				1	1	PNG, SEA
Q3				1	1	PNG, SEA
R*				2	2	
R11b				2	2	MSEA
R21	3	3		1	7	Peninsular Malaysia
R22a				2	2	MSEA
R22b				3	3	Peninsular Malaysia
R22c				1	1	MSEA
R6a1b				1	1	India, MSEA
R7a1				1	1	South Asia, Peninsular Malaysia, Brazil
R9b1a1a	2		3	1	6	SEA
R9b2				1	1	MSEA
U1a3				1	1	Peninsular Malaysia
U2b1				1	1	Peninsular Malaysia
U7a3				2	2	Peninsular Malaysia
Y2a1				2	2	ISEA, Peninsular Malaysia
Total	19	8	13	186	226	

## Appendix D

List of the mtDNA genomes completely sequenced in the present study.

ID	Haplogroup	Variants	Location	Ethnic/Sub-location
BG101	A5b	73 152 235 263 315.1C 522-523d 663 750 961 965.2C 1438 1709 1736 2706 4248 4769 4824 5054C 7028 8260 8563 8794 8860 11536 11719 12705 14766 15326 15442 16126 16223 16290 16319 16519	Pontian, Johor	Bugis
BCK03	B4a1a	73 146 263 292 309.2C 315.1C 522-523d 750 1438 2706 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14766 15326 15746 16086 16182C 16183C 16189 16217 16261 16519	Bachuk, Kelantan	Bachok
BCK04	B4a1a	73 146 263 309.2C 315.1C 522-523d 750 1438 2706 4209 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14766 15326 15746 16181 16182C 16183C 16189 16217 16261 16519	Bachuk, Kelantan	Bachok
BJ153	B4a1a	73 146 263 309.2C 315.1C 522-523d 750 1438 2706 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14766 15326 15746 16178 16182C 16183C 16189 16217 16261 16519	Parit Buntar, Perak	Banjar
BJ156	B4a1a	73 146 263 309.2C 315.1C 522-523d 750 1438 2706 3777 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11339 11719 12239 14766 15326 15746 16182C 16183C 16189 16217 16261 16399 16519	Parit Buntar, Perak	Banjar
MI36	B4a1a	73 146 263 309.2C 315.1C 522-523d 750 1438 2706 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14587 14766 15326 15746 16182C 16183C 16189 16217 16261 16519	Sri Menanti, Negeri Sembilan	Minangkabau
RW162	B4a1a	73 146 263 309.2C 315.1C 522-523d 750 1438 2706 3606 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14766 15326 15746 16182C 16183C 16189 16217 16261 16519	Gopeng, Perak	Rawa
RW174	B4a1a	73 146 263 309.1C 315.1C 522-523d 750 1438 2706 4769 5072 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 13260 14766 15326 15746 16182C 16183C 16189 16217 16230R 16261 16519	Gopeng, Perak	Rawa
BJ119	B4a1a1a	73 146 263 315.1C 522-523d 750 1438 2706 4769 5465 6719 7028 8281.9BPd 8860 9123 10238 11719 12239 14022 14766 15326 15746 16182C 16183C 16189 16217 16247 16261 16519	Kuala Kurau, Perak	Banjar
105A	B4a1c	73 263 310 315.1C 477 522-523d 709 750 1438 2706 4769 5147 5465 7028 7262 8281.9BPd 8860 9123 10238 11719 12192 14766 15016 15326 15784 16129 16182C 16183C 16189 16217 16261 16519	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai



KDH03	B4a1c	73 146 263 309.2C 315.1C 522-523d 709 750 1438 2706 4769 5094 5465 7028 8281.9BPd 8860 9123 10238 11719 13182 14766 15326 16182C 16183C 16189 16217 16519	Lembah Bujang, Kedah	Lembah Bujang
MB10	B4a1c4	73 146 263 315.1C 522-523d 709 750 1438 2706 4769 5465 7028 8281.9BPd 8860 9123 10238 10907 11719 12904 14766 15326 16129 16182C 16183C 16189 16217 16261 16519	Kota Bahru, Kelantan	Kota Bahru
BJ135	B4b1a2	73 204 207 263 315.1C 499 750 827 1438 2706 3308 4769 4820 6023 6216 6413 7028 8281.9BPd 8860 11719 13590 14766 15326 15535 16093 16136 16182C 16183C 16189 16194C 16195 16217 16519	Kuala Kurau, Perak	Banjar
BG094	B4c1b2a2	73 146 150 195 263 309.1C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 7079 8281.9BPd 8772 8860 11719 14766 15301 15326 16346 16111G 16140 16182C 16183C 16189 16217 16274 16335 16519	Pontian, Johor	Bugis
BJ133	B4c1b2a2	73 146 150 195 263 310 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11719 14766 15301 15326 15346 16140 16182C 16183C 16189 16217 16274 16278 16335 16519	Kuala Kurau, Perak	Banjar
JW62	B4c1b2a2	73 146 150 152 195 263 309.1C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 10160 11719 11887 14766 15301 15326 15346 15943G 16140 16182C 16183C 16189 16217 16274 16519	Semerah, Johor	Jawa
MB34	B4c1b2a2	73 146 150 195 263 309.1C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11719 14766 15301 15326 15346 16140 16182C 16183C 16189 16217 16220 16274 16335 16519	Kota Bahru, Kelantan	Kota Bahru
MI28	B4c1b2a2	73 146 150 195 263 309.2C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11719 14766 15301 15326 15346 16140 16182C 16183C 16189 16217 16274 16335 16519	Sri Menanti, Negeri Sembilan	Minangkabau
MI30	B4c1b2a2	73 146 150 195 263 309.2C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11719 14766 15301 15326 15346 16140 16154 16182C 16183C 16189 16217 16274 16335 16519	Sri Menanti, Negeri Sembilan	Minangkabau
MI50	B4c1b2a2	73 146 150 195 263 309.1C 315.1C 709 750 1119 1438 2706 3497 3571 3666 4769 7028 8281.9BPd 8772 8860 11719 14180 14766 15301 15326 15346 15884 16140 16182C 16183C 16189 16217 16274 16335 16519	Lenggeng, Negeri Sembilan	Minangkabau
MI61	B4c1b2a2	73 146 150 195 263 309.1C 315.1C 709 750 1119 1438 1719 2706 3497 3571 3666 4769 7028 8281.9BPd 8772 8860 11719 14766 15301 15326 15346 15884 16140 16182C 16183C 16189 16217 16240 16274 16335 16519	Lenggeng, Negeri Sembilan	Minangkabau
RP02	B4c1b2a2	73 146 150 195 263 309.2C 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11150 11719 13720 14452 14766 15301 15326 15346 16140 16182C 16183C 16189 16213 16217 16274 16335 16519	Rantau Panjang, Kelantan	Rantau Panjang

RW179	B4c1b2a2	16T 73 146 150 195 263 315.1C 709 750 1119 1438 2706 3497 3571 4769 7028 8281.9BPd 8772 8860 11368 11719 14766 15172 15301 15326 15346 16140 16182C 16183C 16189 16217 16274 16311 16335 16519	Gopeng, Perak	Rawa
KT36	B4c2	73 263 315.1C 750 1119 1438 2706 4769 5108 5471 7028 8281.9BPd 8860 11719 14088 14209 14319 14766 15326 15346 16147 16183C 16184A 16189 16192 16217 16235 16519	Pengkalan Hulu, Perak	Semang Kintak
KT43	B4c2	73 263 315.1C 750 1119 1438 2706 4769 5108 5471 7028 8281.9BPd 8860 11719 14088 14209 14319 14766 15326 15346 16147 16183C 16184A 16189 16192 16217 16235 16519	Pengkalan Hulu, Perak	Semang Kintak
BJ126	B4c2	73 263 309.1C 315.1C 750 1119 1438 2706 4769 5108 6221 7028 8281.9BPd 8860 11719 14088 14209 14766 15326 15346 16147 16183 16184A 16189 16217 16235 16519	Kuala Kurau, Perak	Banjar
BJ127	B4c2	73 263 309.2C 315.1C 750 1119 1438 2706 4769 5108 7028 8281.9BPd 8860 11719 14088 14209 14766 14793 15326 15346 16147 16183C 16184A 16189 16217 16235 16235 16519	Kuala Kurau, Perak	Banjar
KS05	B4c2b	73 263 309.2C 315.1C 750 1119 1438 2706 4769 5108 6221 7028 8281.9BPd 8654 8860 11719 14088 14209 14766 15326 15346 16147 16183C 16184A 16189 16217 16235 16519	Baling, Kedah	Senoi Semai
BJ132	B4c2b	73 204 263 309.1C 315.1C 750 1119 1438 2220.1T 2706 4769 5108 5752d 6221 7028 8281.9BPd 8860 11719 14088 14209 14766 15326 15346 16147 16183C 16184A 16189 16217 16235 16261 16356 16519	Kuala Kurau, Perak	Banjar
BJ154	B4c2b	73 263 309.2C 315.1C 750 1119 1438 2706 4769 5108 6221 7028 7501 8281.9BPd 8860 11719 14088 14209 14766 15326 15346 16147 16183C 16184A 16189 16217 16235 16519	Parit Buntar, Perak	Banjar
60B	B5a1a	73 210 263 309.1C 315.1C 522-523d 709 750 1438 2706 3537 4562 4769 6960 7028 8281.9BPd 8584 8860 9670 9950 10398 10915 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16266A 16519	Jeli, Kelantan	Semang Batek
BCK10	B5a1a	73 210 263 309.2C 315.1C 522-523d 709 750 1438 2706 3537 4769 6746 6960 7028 8134 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16167 16183C 16189 16266A 16519	Bachuk, Kelantan	Bachok
BJ134	B5a1a	73 210 263 309.1C 315.1C 522-523d 709 750 1438 2706 3537 4769 5894 6960 7028 8281.9BPd 8584 8860 9950 10398 11719 13145 14766 15235 15326 16140 16183C 16189 16266A 16519	Kuala Kurau, Perak	Banjar
BJ142	B5a1a	73 210 263 309.2C 315.1C 522-523d 709 750 1438 2706 3537 4769 5894 6960 7028 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16266A 16519	Parit Buntar, Perak	Banjar

JW85	B5a1a	73 210 263 315.1C 522-523d 709 750 1438 2706 3537 4769 6340 6960 7028 8281.9BPd 8584 8680 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16266A 16293 16519	Muar, Johor	Jawa
JW89	B5a1a	73 210 263 315.1C 522-523d 709 750 1438 2706 3537 4769 6368G 6960 7028 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16169A 16183C 16189 16209 16266A 16519	Muar, Johor	Jawa
MB31	B5a1a	73 210 263 309.1C 315.1C 522-523d 709 750 1438 2706 3537 4769 5821 6960 7028 8020 8281.9BPd 8584 8860 9950 10398 11446 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16207 16266A 16311 16362 16519	Kota Bahru, Kelantan	Kota Bahru
MC03	B5a1a	73 210 263 315.1C 503 522-523d 709 750 1438 2706 3537 4769 7028 7394 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16266A 16293 16519	Machang, Kelantan	Machang
RP14	B5a1a	73 210 263 309.1C 315.1C 522-523d 709 723 750 1438 2706 3537 4769 6960 7028 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 16140 16183C 16189 16266A 16519	Rantau Panjang, Kelantan	Rantau Panjang
RP31	B5a1a	73 210 263 309.1C 315.1C 522-523d 709 750 1438 2706 3537 4769 6960 7028 8134 8281.9BPd 8584 8860 9950 10398 11719 13145 13395 14766 15235 15326 15628 16140 16167 16183C 16189 16266A 16519	Rantau Panjang, Kelantan	Rantau Panjang
RW164	B5a1b	73 210 263 315.1C 522-523d 709 750 1438 2706 3537 4769 6960 7028 7852 7864 8281.9BPd 8584 8860 9950 10398 10754 11719 14766 14989 15235 15326 16140 16183C 16189 16266A 16519	Gopeng, Perak	Rawa
BJ145	B5a1d	73 152 210 263 309.2C 315.1C 522-523d 709 750 1438 2706 3537 4086 4769 6960 7028 8281.9BPd 8584 8860 9950 10398 11465 11662 11719 14766 15235 15326 16140 16182C 16183C 16189 16214 16223 16261 16266A 16270 16519	Parit Buntar, Perak	Banjar
MC14	B5a1d	73 152 210 263 315.1C 522-523d 709 750 1438 2706 3537 4086 4769 6960 7028 8281.9BPd 8584 8860 9950 10398 11465 11608 11719 14766 15235 15326 16140 16182C 16183C 16189 16261 16266A 16519	Machang, Kelantan	Machang
JW83	B5b1c	10 73 103 152 204 263 315.1C 522-523d 709 750 960.1C 1438 1598 2706 3480 3565 3819 4769 5836 7028 7771 8281.9BPd 8467 8584 8784 8829 8860 8943 9950 10274 10398 11506 11719 12361 12858 14766 15223 15326 15508 15662 15851 15927 16067 16140 16183C 16189 16243 16519	Muar, Johor	Jawa

RW161	B5b1c	73 103 152 263 315.1C 522-523d 709 750 960.1C 1438 1598 2706 3480 3819 4769 5836 7028 7771 8281.9BPd 8467 8584 8784 8829 8860 9950 10274 10398 11719 12361 14766 15223 15326 15508 15662 15851 15927 16140 16183C 16189 16243 16519	Gopeng, Perak	Rawa
8A	B5b1c	73 103 152 204 263 309.1C 315.1C 522-523d 709 750 960.1C 1438 1598 2706 3480 3819 4769 5836 7028 7771 8281.9BPd 8467 8584 8784 8829 8860 9950 10274 10398 11719 12361 14766 15223 15326 15508 15662 15851 15927 16140 16183C 16189 16243 16294 16354 16519	Semang Batek	Semang Batek
137A	B6a1a	73 263 309.1C 315.1C 356.1C 750 1438 1719 2706 4093 4769 5894C 6758 7028 8281.9BPd 8860 9452 11719 11914 12950 13928C 14305 14766 15326 16051 16183C 16189 16519 16527	Pilah Temuan, Negeri Sembilan	Aboriginal Malay Temuan
AC11	B6a1a	73 263 315.1C 356.1C 750 1187 1438 1719 2706 4093 4615 4769 5893.2C 5894C 6758 7028 8281.9BPd 8860 9452 11719 11914 12950 13928C 14305 14766 15326 16051 16183C 16189 16194C 16195 16239 16519 16527	Yan, Kedah	Acheh
JW69	B6a1a	73 195 263 309.1C 315.1C 356.1C 750 1438 1719 2706 4047 4093 4769 5894C 6758 7028 8281.9BPd 8860 9452 11719 11914 12141 12950 13928C 14305 14766 15326 16051 16183C 16189 16266A 16519	Semerah, Johor	Jawa
MI49	B6a1a	73 263 315.1C 356.1C 750 1187 1438 1719 2706 4093 4615 4769 5893.2C 5894C 6758 7028 8281.9BPd 8860 9452 11719 11914 12950 13928C 14305 14766 15326 16051 16183C 16189 16194C 16195 16519 16527	Sri Menanti, Negeri Sembilan	Minangkabau
MB02	C7a	44.1C 73 249d 263 315.1C 489 750 1438 2706 3552A 4715 4769 5821 6338 7028 7196A 7853 8584 8701 8860 9540 9545 10398 10400 10873 11719 11914 12705 13263 14318 14766 14783 15043 15301 15326 15487T 16223 16298 16327 16519	Kota Bahru, Kelantan	Kota Bahru
KDH14	C7a1d	73 146 249d 263 297 315.1C 489 750 1438 2706 2905 3552A 4715 4769 5821 6338 7028 7196A 7853 8584 8701 8860 9540 9545 10398 10400 10873 11719 11914 12705 12957 13263 13879A 14318 14766 14783 14978 15043 15301 15326 15487T 16086 16223 16242 16256 16298 16327 16519	Lembah Bujang, Kedah	Lembah Bujang
RP05	E1a1a1	73 146 150 263 315.1C 489 522-523d 750 1438 2706 3027 3197 3229.1A 3705 4248 4491 4769 6023 6620 7028 7598 8701 8843 8860 9540 10398 10400 10834 10873 11719 12705 13254 13626 14577 14783 15043 15301 15326 16223 16288 16291 16362 16390 16519	Rantau Panjang, Kelantan	Rantau Panjang
KDH06	F1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 1860 2706 3970 4086 4769 6392 6962 7028 8589 8860 9053 9548 10310 10609 11719 12406 12634 12882 13759 13928C 14766 15326 15884 16129 16162 16172 16304 16519	Lembah Bujang, Kedah	Lembah Bujang

MB25	F1a1	73 249d 251 263 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8860 9053 9548 10310 10463 10609 11719 12406 12630 12882 13759 13928C 14766 15326 16129 16162 16172 16304 16335 16519	Kota Bahru, Kelantan	Kota Bahru
20B	F1a1a	73 152 249d 263 315.1C 522-523d d 750 1438 2706 3970 4086 4769 6040 6392 6962 7028 8149 8860 9053 9548 10310 10609 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16304 16519	Kuala Betis, Kelantan	Senoi Temiar
BCK02	F1a1a	73 249d 263 309.1C 315.1C 522-523d 750 1438 2483 2706 3777 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16304 16519	Bachuk, Kelantan	Bachok
BJ125	F1a1a	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8281.9BPd 8860 9053 9548 10310 10609 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16304 16368 16519	Kuala Kurau, Perak	Banjar
JW66	F1a1a	73 249d 263 315.1C 522-523d 750 1438 2706 3777 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11719 11923 12406 12882 13759 13928C 14470 14766 15326 15773 16108 16129 16162 16172 16304 16519	Semerah, Johor	Jawa
KDH16	F1a1a	73 249d 263 309.1C 315.1C 750 1438 2706 3777 3970 4086 4534 4769 6392 6962 7028 8149 8658 8860 9053 9548 10310 10609 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16304 16519	Lembah Bujang, Kedah	Lembah Bujang
KDH26	F1a1a	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9260 9468 9548 10310 10609 11719 12406 12882 13759 13928C 14766 15326 16092 16108 16129 16162 16172 16234 16299 16304 16519	Lembah Bujang, Kedah	Lembah Bujang
TPT01	F1a1a	73 249d 263 309.1C 315.1C 438 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11719 12406 12882 13329 13759 13928C 14766 15326 16108 16129 16162 16172 16183C 16189 16234 16519	Tumpat, Kelantan	Tumpat
136A	F1a1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16147 16162 16172 16304 16519	Pilah, Negeri Sembilan	Aboriginal Malay Jakun
159B	F1a1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6797 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12406 12820 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16274 16519	Gombak, Selangor	Senoi Temiar
LN12	F1a1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12406 12820 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16519	Lenggong, Perak	Senoi Temiar

LN13	F1a1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12405 12406 12820 12882 13759 13928C 14766 15326 16108 16129 16172	Lenggong, Perak	Senoi Temiar
BJ143	F1a1a1	73 249d 263 315.1C 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10463 10609 11215 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16294 16304 16362 16519	Parit Buntar, Perak	Banjar
JW71	F1a1a1	73 249d 263 309.1C 315.1c 522-523d 750 1438 2706 3970 4086 4769 6253 6392 6962 7028 8149 8860 9053 9377 9548 10310 10609 11215 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16188 16304 16519	Semerah, Johor	Jawa
KDH08	F1a1a1	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 7270 8149 8860 9053 9548 10187 10310 10609 11215 11719 12406 12882 13759 13812 13928C 14766 15326 16108 16129 16162 16172 16239 16304 16327 16519	Lembah Bujang, Kedah	Lembah Bujang
RP18	F1a1a1	73 249d 263 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6917 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12406 12882 13153 13759 13928C 14766 15326 16108 16129 16147 16162 16172 16304 16519	Rantau Panjang, Kelantan	Rantau Panjang
RP20	F1a1a1	73 249d 263 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8149 8860 9053 9548 10310 10609 11215 11719 12406 12882 13759 13928C 14766 15326 16108 16129 16162 16172 16304 16357 16519	Rantau Panjang, Kelantan	Rantau Panjang
AC01	F1a1c	73 249d 263 315.1C 522-523d 548 750 1438 2706 3970 4086 4769 6392 6962 7028 8860 9053 9548 10211 10310 10609 11593 11719 12406 12882 13135 13759 13928C 14766 15326 16129 16162 16172 16224 16304 16519	Yan, Kedah	Acheh
BCK09	F1a1d	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8860 9053 9548 10310 10609 11380 11719 12406 12882 13753 13759 13928C 14766 15326 16129 16162 16172 16304 16399 16519	Bachuk, Kelantan	Bachok
BJ131	F1a3a	73 236 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4086 4769 6392 6962 7028 8860 9053 9554 9944 10310 10609 11719 11899 12406 12882 13748 13759 13928C 14233 14766 15326 15565 16129 16172 16304 16311 16519	Kuala Kurau, Perak	Banjar
MB39	F1a4a	73 152 249d 263 309.1C 315.1C 520d 521d 522-523d 750 1438 2706 3970 4086 4769 5985 6392 6962 7028 8277 8860 8998 9053 9548 10310 10609 11719 12406 12882 13422 13759 13928C 14766 15326 15445 16129 16172 16294 16304 16362 16519	Kota Bahru, Kelantan	Kota Bahru
BJ147	F1f	73 249d 263 315.1C 522-523d 750 1438 1457 2706 3970 4715 4769 5147 6353 6392 6515 6962 7028 8860 9053 10310 10609 11719 12406 12771 12882 13759 13928C 14766 15326 16129 16172 16183C 16189 16304 16519	Parit Buntar, Perak	Banjar

MB09	F1f	73 249d 263 315.1C 522-523d 750 1438 2706 3970 4715 4769 6392 6515 6962 7028 8860 9053 10310 10609 11719 12406 12771 12882 13759 13928C 14766 15326 16129 16172 16304 16519	Kota Bahru, Kelantan	Kota Bahru
MC24	F1f	73 249d 263 279 309.1C 315.1C 522-523d 750 1438 2392 2706 3970 4715 4769 6392 6515 6962 7028 8860 9053 10310 10609 11719 12151 12406 12771 12882 13759 13928C 14766 15326 16129 16172 16291 16304 16519	Machang, Kelantan	Machang
MI56	F1f	73 249d 263 309.1C 315.1C 522-523d 750 1438 2706 3970 4715 4769 6392 6515 6962 7028 8860 9053 10310 10609 11719 12406 12771 12882 13759 13928C 14766 15326 16129 16172 16301 16304 16400 16519	Lenggeng, Negeri Sembilan	Minangkabau
KDH05	F3a1	73 207 249d 263 309.1C 315.1C 709 750 1438 2706 3434 3970 4769 4824 4991 5585 5894 5913 5978 6392 7028 8860 10310 10320 11065 11719 12621 13928C 14766 14971 15326 15412G 16260 16298 16355 16362	Lembah Bujang, Kedah	Lembah Bujang
MB33	F3a2	73 152 195 207 249d 263 309.1C 315.1C 750 1438 2706 3390 3434 3970 4769 5585 5913 5978 6392 7028 7094 8860 10310 10320 11065 11719 12237 12621 13928C 14766 15326 16209 16298 16355 16362	Kota Bahru, Kelantan	Kota Bahru
MC01	F3b1a	73 150 152 249d 263 309.1C 315.1C 750 1438 2706 3434 3970 4769 5076 5585 5913 5978 6392 6791 7028 8838 8860 9947 10310 10320 11209 11719 13928C 14766 15326 15784 16220C 16265 16298 16311 16362	Machang, Kelantan	Machang
RW172	F4b	73 249d 263 309.1C 315.1C 573.3C 750 1438 2706 3970 4769 5263 6392 6653 7028 8020 8575 8603 8860 10097C 10310 11719 11908 12630 13928C 14766 15326 15670 16170 16218 16304 16311 16526	Gopeng, Perak	Rawa
BCK06	M12	73 263 297 315.1C 489 522-523d 750 960.1C 1438 2706 3579 4170 4769 5036 5580 7028 8251 8701 8781 8860 9540 10398 10400 10873 11569 11719 12030 12372 12705 13242 14364 14569 14727 14766 14783 15010 15043 15301 15326 16223 16234 16239 16290 16309 16362 16391	Bachuk, Kelantan	Bachok
RP21	M13'46'61	73 146 152 195 263 315.1C 489 522-523d 750 951 1438 2706 4226 4769 5262 5773 7028 7040 7232 8701 8860 9540 10398 10400 10873 11719 12705 13359 14766 14783 15043 15301 15326 15601 16223 16311 16362 16519	Rantau Panjang, Kelantan	Rantau Panjang
KT05	M17a1a	64 73 150 263 309.1C 315.1C 489 522-523d 750 862 930C 1438 2706 4769 7028 7170 8251 8701 8860 9540 10324 10398 10400 10873 11016 11719 11908 12705 12711 12804 12973 14716 14766 14783 15043 15301 15326 15530 15802 15941 16093 16129 16209 16223 16261 16278 16325 16519	Pengkalan Hulu, Perak	Semang Kensiu
KT18	M17a1a	64 73 150 263 309.1C 315.1C 489 522-523d 750 862 930C 1438 2706 4769 7028 7170 8251 8701 8860 9540 10202 10398 10400 10873 11016 11719 11908 12705 12711 12804 12973 14716 14766 14783 15043 15301 15326 15530 15802 15941 16093 16129 16209 16223 16261 16278 16325 16519	Pengkalan Hulu, Perak	Semang Kensiu

MB27	M17c	73 152 259 263 315.1C 489 522-523d 750 930C 1438 1598 2706 3316 3766 4316 4769 5074 5822 7028 7346 8701 8860 9377 9540 10398 10400 10873 11002 11150 11253 11719 12705 12973 13767 13830 14766 14783 15043 15301 15326 16111A 16150 16209 16223 16304	Kota Bahru, Kelantan	Kota Bahru
MC06	M17c	73 263 315.1C 489 522-523d 709 750 930C 1438 1598 1943 2706 4769 4917 7028 8701 8860 9540 10103 10166 10398 10400 10873 11204 11719 12012 12705 12973 13651 14766 14783 15043 15301 15326 15853 16209 16223 16233 16274 16304	Machang, Kelantan	Machang
BCK08	M20	73 150 152 249d 263 315.1C 316 455.1T 489 750 1438 2706 2963 4697A 4769 4772 6915 7028 8639G 8701 8853 8860 9540 10253 10398 10400 10873 11719 11914 12354 12705 13708 14110 14766 14783 15043 15301 15326 16223 16272	Bachuk, Kelantan	Bachok
RP33	M20a	73 199 249d 263 309.2C 315.1C 316 489 522-523d 750 1438 2706 3200 3714 4385T 4769 4772 7028 7433 8701 8853 8860 9127 9254 9512 9540 10274 10398 10400 10873 11350 11719 11914 12354 12705 14082 14110 14766 14783 15043 15226C 15301 15326 15497 15691 16129 16209 16223 16272 16311 16519	Rantau Panjang, Kelantan	Rantau Panjang
MB03	M20a1a	73 146 225 249d 263 315.1C 316 489 522-523d 750 1438 2706 3200 3714 4385T 4769 4772 7028 7433 8701 8853 8860 9127 9512 9540 10274 10398 10400 10679 10873 11150 11719 11914 12354 12705 14110 14766 14783 14974 15043 15301 15326 15691 16129 16209 16223 16272 16519	Kota Bahru, Kelantan	Kota Bahru
JW74	M20a1a1	73 143 152 225 249d 263 315.1C 316 489 520d 521d 522- 523d 750 1438 2706 3200 3714 4385T 4769 4772 7028 7433 8701 8853 8860 9127 9512 9540 10274 10398 10400 10679 10873 11719 11914 12354 12705 14110 14766 14783 14974 15043 15301 15326 15691 16086 16129 16209 16223 16272 16519 16527	Semerah, Johor	Jawa
KDH17	M20a1a1	73 143 152 225 249d 263 315.1C 316 489 522-523d 750 1438 2706 3200 3714 4385T 4769 4772 7028 7433 8701 8853 8860 8974 9127 9512 9540 10274 10398 10400 10679 10873 11719 11914 12354 12705 13659 14110 14766 14783 14974 15043 15301 15326 15691 16086 16129 16209 16223 16272 16519	Lembah Bujang, Kedah	Lembah Bujang
MB24	M20a1a1	73 152 225 249d 263 315.1C 316 489 520d 521d 522-523d 750 1438 2706 3200 3714 4385T 4769 4772 7028 7433 8701 8853 8860 9127 9512 9540 10274 10325 10398 10400 10679 10873 11719 11914 12354 12705 14110 14766 14783 14974 15043 15301 15326 15691 16086 16129 16209 16223 16266 16272 16519	Kota Bahru, Kelantan	Kota Bahru



JW82	M20a1a2	73 152 225 249d 263 309.1C 315.1C 316 489 522-523d 750 956 1438 2706 3200 3714 4385T 4769 4772 5529T 7028 7433 7912 8701 8853 8860 9127 9512 9540 10274 10398 10400 10679 10873 11719 11914 12354 12705 14110 14766 14783 14974 15043 15301 15326 15691 16129 16209 16223 16249 16272 16519	Muar, Johor	Jawa
RP07	M21c1	73 152 195 263 309.1C 315.1C 489 522-523d 750 1438 2706 3915 4769 5108 6287 7028 7765 7861 8293 8701 8860 9116 9540 10398 10400 10873 11482 11719 12705 12940 13590 14766 14783 14809 15043 15301 15326 16093 16223 16519	Rantau Panjang, Kelantan	Rantau Panjang
BJ137	M21c2	73 146 263 315.1C 489 522-523d 750 1415 1438 1808 2706 4769 5054 5108 5186 6923 7028 7861 8701 8860 9210 9449 9540 10057 10302 10398 10400 10873 11482 11719 12612 12705 13105 13329 14766 14783 15043 15250 15301 15326 15955.1A 16093 16223 16274 16278 16519	Kuala Kurau, Perak	Banjar
KDH13	M26b	73 146 249 263 315.1C 489 522-523d 750 1438 1719 2706 3531 4021 4769 4901 6026T 7028 7289 8563 8701 8860 8970 9380 9540 10256 10398 10400 10873 11140 11425 11719 12705 13708 14040 14766 14783 15043 15301 15326 15514 16111 16129 16140 16172 16189 16194C 16223 16278	Lembah Bujang, Kedah	Lembah Bujang
JW91	M26b2	73 249 263 309.1C 315.1C 319 489 522-523d 750 1438 2706 4021 4769 4901 7028 8701 8860 8970 9380 9540 10256 10398 10400 10873 11140 11719 12705 12741 13708 14040 14766 14783 15043 15301 15326 15574 15850 16092 16093 16140 16169 16172 16189 16223 16278	Muar, Johor	Jawa
MB18	M30a1	73 150 195A 263 315.1C 489 513 522-523d 750 1438 2162 2706 4216 4769 6366 7028 8701 8860 9540 10398 10400 10873 11719 11928 12007 12211 12705 14766 14783 15043 15301 15326 15431 16223	Kota Bahru, Kelantan	Kota Bahru
MI37	M37e	73 263 315.1C 489 522-523d 551 750 1438 2706 4769 7028 8701 8860 9509 9540 10398 10400 10556 10873 11050 11719 11974 12007 12705 14766 14783 15043 15262 15301 15326 16111 16184 16189 16223 16295 16296 16311 16519	Sri Menanti, Negeri Sembilan	Minangkabau
MC05	M50a1	73 146 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3316 4769 7028 7226 7844 8701 8860 9540 9608 9629 10398 10400 10538A 10873 11365 11719 12705 14182 14766 14783 15043 15301 15326 15616 15663 16209 16223 16224 16263 16278 16319	Machang, Kelantan	Machang
RP09	M50a1a	73 146 150 151 263 309.2C 315.1C 489 522-523d 750 1438 2706 3316 4769 7028 7226 8701 8860 9540 9608 10398 10400 10538A 10873 11365 11719 12705 14182 14766 14783 15043 15301 15326 15616 15663 16093 16209 16223 16224 16263 16278 16319	Rantau Panjang, Kelantan	Rantau Panjang

MB44	M72a1a	73 263 309.1C 315.1C 489 750 1438 1872 2706 4058 4769 7028 8701 8860 9540 10398 10400 10873 11719 12705 12753 13281 14233 14766 14783 15043 15301 15326 15644 15820 15932 16124 16166d 16175 16214 16223 16263 16519	Kota Bahru, Kelantan	Kota Bahru
MI35	M72a1b	73 263 309.1C 315.1C 489 573.2C 750 1438 1872 2706 4769 7028 8701 8860 9540 10398 10400 10873 11719 12705 12753 13176 14233 14766 14783 15043 15301 15326 15644 15820 16124 16166d 16214 16223 16362	Sri Menanti, Negeri Sembilan	Minangkabau
RP10	M77	73 194 263 309.1C 315.1C 489 522-523d 750 1409d 1438 2706 4065 4769 7028 8419 8701 8860 9540 10398 10400 10873 11719 12705 13105 13407 13542 14178 14544 14766 14783 15043 15301 15326 16093 16129 16189 16213 16218 16223 16519	Rantau Panjang, Kelantan	Rantau Panjang
100B	M7c3c	73 146 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11335 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16223 16295 16362 16519	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai
BJ136	M7c3c	73 146 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16223 16295 16362 16519	Kuala Kurau, Perak	Banjar
JW73	M7c3c	73 146 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5442 6455 7028 8701 8860 9389 9540 9824 10398 10400 10873 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16223 16362 16519	Semerah, Johor	Jawa
JW78	M7c3c	73 150 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5267 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16187 16223 16295 16362 16519	Muar, Johor	Jawa
MB15	M7c3c	73 146 199 263 309.1C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16223 16295 16355 16362 16519 16524	Kota Bahru, Kelantan	Kota Bahru
RP04	M7c3c	73 146 199 204 263 309.2C 315.1C 489 522-523d 750 1438 1819 2706 3606 4071 4769 4820 4850 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11665 11719 12091 12705 14766 14783 15043 15236 15301 15326 16093 16223 16295 16362 16519	Rantau Panjang, Kelantan	Rantau Panjang
RP26	M7c3c	73 146 199 263 309.2C 315.1C 489 522-523d 750 1438 2706 3606 4071 4769 4850 5442 6455 7028 8701 8860 9540 9824 10398 10400 10873 11665 11719 11914 12091 12705 13884 14766 14783 15043 15236 15301 15326 16176G 16223 16295 16296 16362 16519	Rantau Panjang, Kelantan	Rantau Panjang

BCK01	N21	73 150 195 263 309.2C 315.1C 337d 532 750 961 1438 2706 4769 6752 7028 8701 8860 9512 10583 11719 12358 12705 13135 13437 14560 14766 15326 16179 16193 16223 16291 16519	Bachuk, Kelantan	Bachok
MI54	N21	73 150 195 263 309.1C 315.1C 337d 750 1438 2706 3552 4769 6752 7028 8701 8860 9512 10583 10752 11368 11719 12705 13135 13437 14560 14766 15326 16193 16291 16519	Lenggeng, Negeri Sembilan	Minangkabau
110B	N21a1a	73 150 195 263 315.1C 337d 750 1438 2706 3552 4769 6752 7028 8701 8860 9512 10583 11719 12705 13135 13437 14560 14766 15326 16193 16291 16327 16519	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai
MC11	R11b	73 185 189 263 315.1C 709 750 1438 2706 4769 5836 7028 8281.9BPd 8860 10031 10398 11061 11719 12950 13269 13681 14322 14766 15326 16086 16182C 16183C 16189 16239 16309 16311 16390 16399 16519	Machang, Kelantan	Machang
MB16	R6a1b	73 215 228 263 315.1C 522-523d 750 1438 2706 4769 7028 8860 8958 11075 11464A 11719 12088 12285 14058 14153 14582 14766 15326 16129 16274 16284 16362 16519	Kota Bahru, Kelantan	Kota Bahru
KT03	R9b1a1a	73 143 146 183 263 309.1C 315.1C 522-523d 573.2C 750 1438 1541 2706 3970 4017 4769 7028 7633 7849 8302 8860 11719 12714 13928C 14766 15326 16093 16192 16288 16304 16309 16390 16519	Pengkalan Hulu, Perak	Semang Jahai
KT38	R9b1a1a	73 143 146 183 263 309.1C 315.1C 522-523d 573.2C 750 1438 1541 2706 3970 4017 4769 7028 7633 7849 8860 11719 12714 13928C 14766 15326 16093 16192 16288 16304 16309 16390 16519	Pengkalan Hulu, Perak	Semang Kintak
100A	R9b1a1a	73 143 152 183 263 309.1C 315.1C 522-523d 573.2C 750 1438 1541 2706 3970 4017 4769 7028 7849 8860 9221 11719 12714 13928C 14766 15326 16086 16170 16223 16288 16304 16309 16390	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai
101A	R9b1a1a	73 143 152 183 263 309.1C 315.1C 522-523d 573.2C 750 1438 1541 2706 3970 4017 4769 7028 7849 8860 9221 11719 12714 13928C 14766 15326 16086 16170 16223 16288 16304 16309 16390	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai
108A	R9b1a1a	73 143 152 183 263 309.1C 315.1C 522-523d 750 1438 1541 2706 3970 4017 4769 7028 7849 8860 9221 11719 12714 13863 13928C 14766 15326 16086 16170 16223 16288 16304 16309 16390	Pos <i>Orang Asli</i>	Aboriginal Malay Semelai
RP15	R9b1a1a	73 143 146 183 263 315.1C 522-523d 573.1C 750 1393 1438 1541 2706 3970 4017 4769 7028 7849 8860 10237 10694 11719 12714 13928C 14766 15326 16192 16288 16304 16309 16390 16519	Rantau Panjang, Kelantan	Rantau Panjang
RW160	U7a3	73 151 152 188 263 309.1C 315.1C 522-523d 750 980 1438 1811 2706 3741 3834 4769 5360 6164 7028 8137 8684 8860 10142 11467 11719 12308 12372 12618 13500 14218 14569 14766 15326 16187 16207 16243 16309 16318T 16519	Gopeng, Perak	Rawa

RW166	U7a3	73 151 152 188 263 309.1C 315.1C 522-523d 750 980 1438 1811 2706 3741 3834 4769 5360 6164 7028 8137 8684 8860 10142 11467 11719 12308 12372 12618 13500 14218 14569 14766 15326 16187 16207 16243 16309 16318T 16519	Gopeng, Perak	Rawa
AC02	Y2a1	73 228 234 263 309.1C 315.1C 482 522-523d 709 750 1438 2706 4769 5147 5417 6791 6941 7028 7859 8392 8567 8860 10398 11299 11719 12161 12705 14178 14693 14766 14914 15244 15301 15326 16126 16231 16311	Yan, Kedah	Acheh

# Appendix E – Haplogroups not found in Malaysia

## Haplogroup M7a

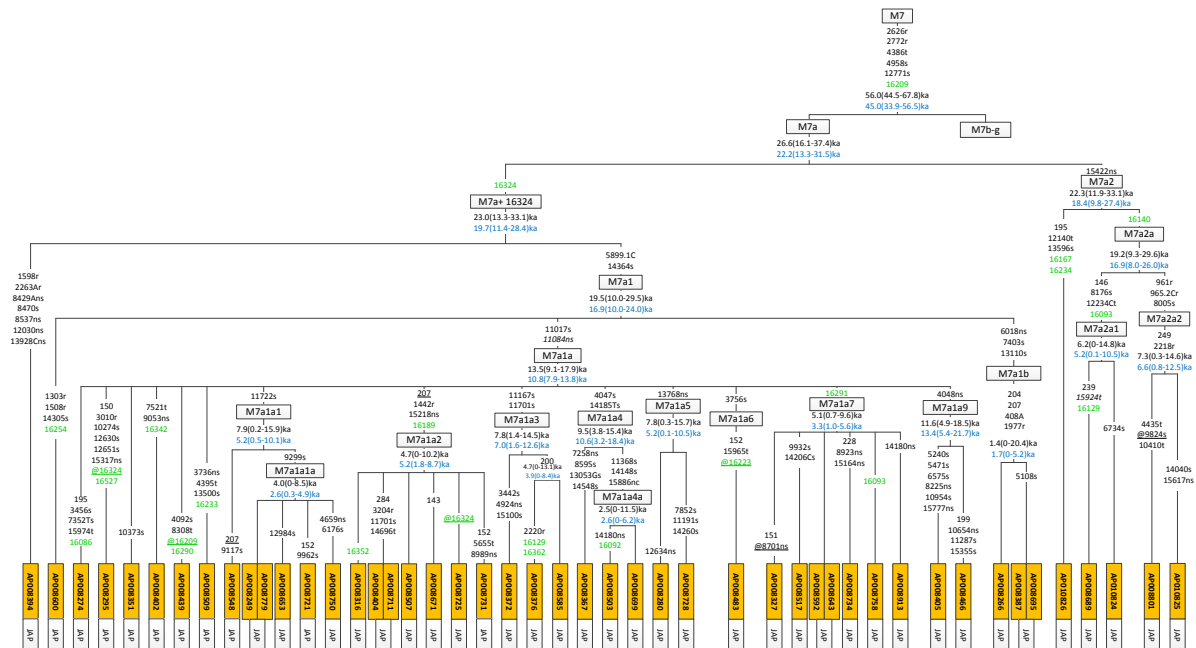
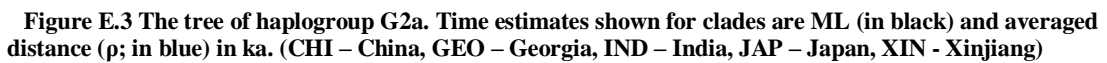


Figure E.1 The tree of haplogroup M7a. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (JAP – Japan)

There are two main clades in M7a: M7a+T16324C (includes M7a1) and M7a2 (Figure E.1). **M7a+T16324C** dates to the LGM at ~23 ka, and **M7a1** ~19.5 ka, which further diverged towards the end of Ice Age into **M7a1a** (~13.5 ka) and **M7a1b** (~1.5 ka). Figure shows that **M7a2** dates to the LGM at ~22 ka and a basal lineage is seen in Shizuoka Japan (Nohira *et al.*, 2010), and its subclades **M7a2a1** (dates to ~6 ka) is seen in Chiba (Nohira *et al.*, 2010) and Aichi (Tanaka *et al.*, 2004), and **M7a2a2** (~7 ka) found in Gunma (Nohira *et al.*, 2010) and Aichi (Tanaka *et al.*, 2004). The whole-mtDNA tree shows M7a has a pre-LGM ancestry in Japan.





**G2b** dates to ~27 ka, and can be divided into G2b1 and G2b2 (Figure E.4). Similar to G2a, G2b is commonly found in China (Kong *et al.*, 2003a; Zheng *et al.*, 2011), Japan (Hartmann *et al.*, 2009; Nohira *et al.*, 2010) and India (Chandrasekar *et al.*, 2009). **G2b1**, dating to ~21 ka, is divided into G2b1a and G2b1b. **G2b1a** dates to ~7 ka and seen in China. **G2b1b** dates to ~9 ka, and a further subclade dates to ~6 ka are restricted to India. **G2b2**

dates to ~23 ka, and is found in Japan (Hartmann *et al.*, 2009; Nohira *et al.*, 2010) and China (Zheng *et al.*, 2011). Nested within is G2b2a (dates to ~18 ka) that is seen in India (Chandrasekar *et al.*, 2009).

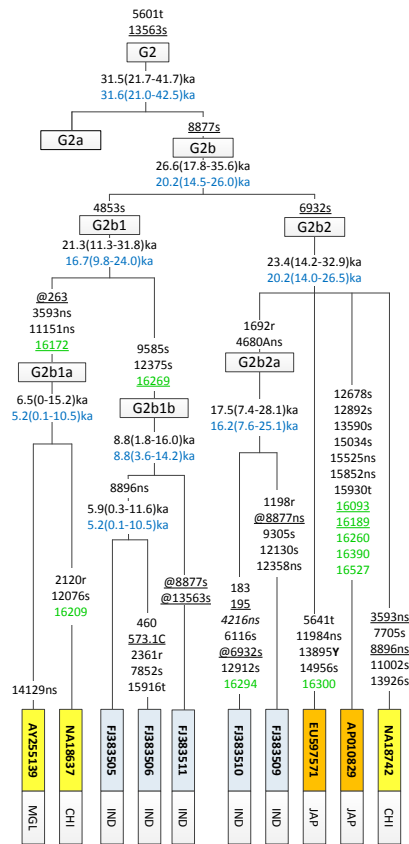


Figure E.4 The tree of haplogroup G2b. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, IND – India, JAP – Japan, MGL – Inner Mongolia, China)

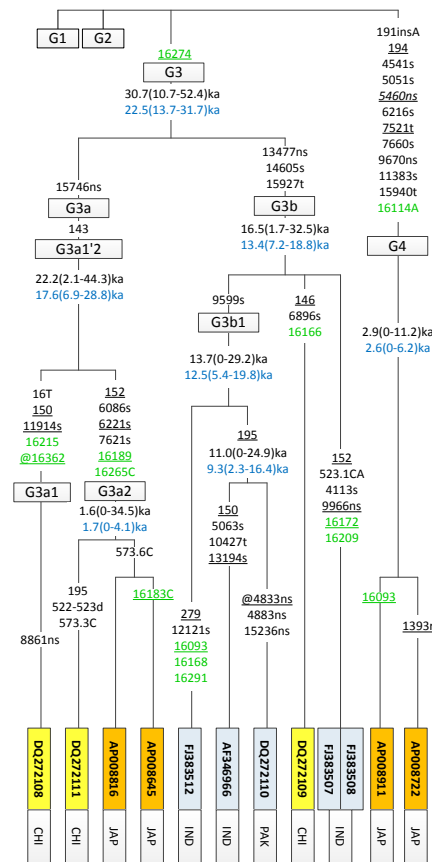
## Haplogroups G3 and G4

Figure E.5 shows **G3** (dates to ~31 ka) which can be divided into G3a1'2 (~22 ka) and G3b (~16.5 ka). **G3a1'2** has two subclades, G3a1 and G3a2, and occurred at lower levels in Japan (Tanaka *et al.*, 2004) and China (Kong *et al.*, 2006). Only one sequence represented G3a1 from China, and hence no age estimations. **G3a2** dates to ~2 ka and seen in China, which spreads into Japan (no dates for subclades defined by indels).

**G3b** dates to ~17 ka with basal lineages seen in China (Kong *et al.*, 2006) and India (Chandrasekar *et al.*, 2009). **G3b1** is restricted to India and Pakistan dating to ~14 ka (Ingman *et al.*, 2000; Kong *et al.*, 2006; Chandrasekar *et al.*, 2009). Although the HVS-I data shows that G3 is predominantly found in Japan (Tanaka *et al.*, 2004), the older lineages on the whole-mtDNA tree are found in China dating to the LGM. Similarly observed in G2b1

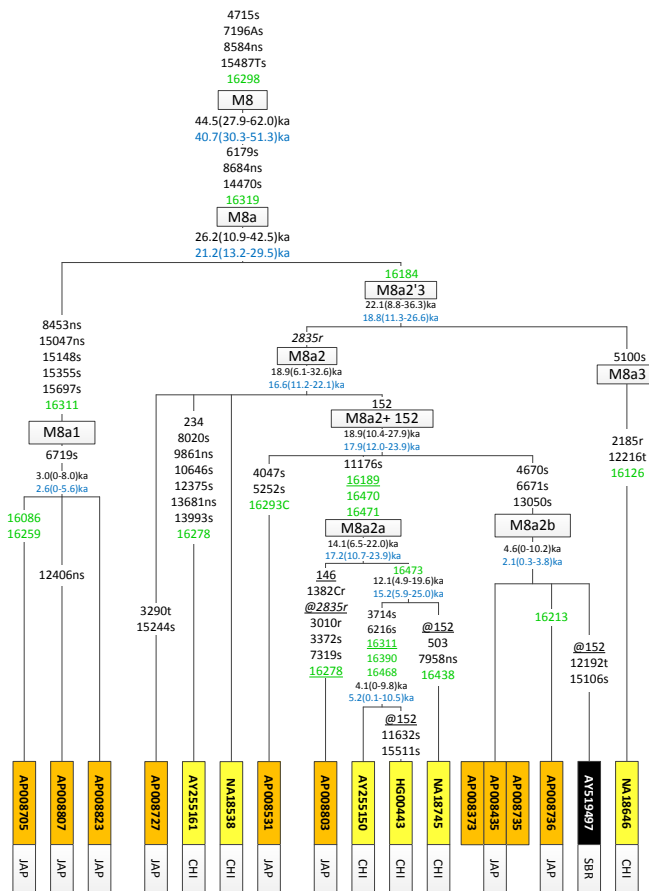


and G2b2, G3b again suggests an ancient haplogroup G population in China before the LGM from which numerous lineages dispersed together into India and Japan around the LGM onwards.



**Figure E.5** The trees of haplogroup G3 and G4. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, IND – India, JAP – Japan, PAK - Pakistan)

## Haplogroup M8a



**Figure E.6** The tree of haplogroup M8a. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, SBR – Siberia, Russia)

**M8a**, dating to ~26 ka, is divided into M8a1 and M8a2'3, and is mainly restricted to China and Japan as mentioned earlier (Figure E.6). **M8a1** has a recent date of ~3 ka and found only in Aichi Japan (Tanaka *et al.*, 2004). **M8a2'3** and M8a2 date to the LGM, ~22 ka and ~19 ka respectively, and are seen in Japan (Tanaka *et al.*, 2004) and China (Kong *et al.*, 2003a; Zheng *et al.*, 2011), with M8a2b (~4.6 ka) also seen in a single Siberian Russian (Starikovskaya *et al.*, 2005). M8a3 is represented by a single instance from China (Zheng *et al.*, 2011).

We can then complement the whole-mtDNA tree with the control-region data included in Tanaka *et al.* (2004; Table 2 & 3 in the paper), where they found M8a at high frequency in South and Eastern China, Taiwan, Korea and Japan. Unfortunately, the frequency table of the HVS-I data were segregated according to different papers instead of by populations/locations, for e.g. population “Ch1” (Table 3 in Tanaka *et al.*, 2004) consists of Lining and Shandong in the north, Yunnan in the south and Changsha in central China, while “Ch2” includes

samples from Xinjiang, Qinghai, Xi'an, Yunnan, Shanghai and Mongolia. Nonetheless, similar to the whole-mtDNA tree, HVS-I data shows M8a is virtually absent in SEA, it is most likely an East Asia haplogroup that pre-dates the LGM.

### Haplogroups C1 and C4

Haplogroup **C1** dates to the LGM ~23 ka (Figure E.7). It is seen in single instances from Japan (Tanaka *et al.*, 2004) and Siberian Russia (Starikovskaya *et al.*, 2005) as **C1a** (~13 ka), and in Native Americans (Ingman *et al.*, 2000; Mishmar *et al.*, 2003) and Mexico (Hartmann *et al.*, 2009) as **C1b** (~20 ka), **C1c1a** and **C1d1** (~1 ka). C1c1a is represented by a single instance from Mexico (Hartmann *et al.*, 2009).

In Figure E.8, **C4**, dating to ~23 ka, with a single basal lineage seen in northern China (Zheng *et al.*, 2011). It is reported to be specific for Altai region in Southern Siberia, although it is also seen in northeastern Asia, East Asia, India as well as Europe (Derenko *et al.*, 2010). **C4a**, dating to ~21 ka, is divided into C4a1'5 and C4a2'3'4. Figure shows **C4a1+16129+195** dates to ~12 ka and subclade **C4a1b** is seen in Xinjiang (Kong *et al.*, 2003a) and southern China (Zheng *et al.*, 2011) dating to ~7 ka. **C4a1c1a** dates to ~7 ka and seen in Evenk, Eskimo (Mishmar *et al.*, 2003) and Siberian Russian (Starikovskaya *et al.*, 2005). **C4a2'3'4** dates to ~20 ka, and represented by a single instance in each subclade; C4a2 is seen in Evenk, Eskimo (Ingman *et al.*, 2000; Starikovskaya *et al.*, 2005) and C4a3 in Beijing, China (Zheng *et al.*, 2011) respectively. Additionally, subclades C4a1b, C4a2a2 and C4a2b (and C7a1a below) are reportedly predominantly Indian haplogroups (Derenko *et al.*, 2010). According to the whole-mtDNA tree published by Derenko *et al.* (2010), the Indian subclades nested within C4a2 that have older dates compared with C4a1 seems to suggest an Indian origin dating to the LGM with east- and northwards late glacial spread into East Asia, Southern Siberia and northeastern Asia.

**C4b** dates to ~12 ka, and includes subclades C4b1, C4b2, C4b3 and C4b8, where the last three are represented by single/similar instance each from Koryak, Eskimo in northeast Siberia, Inner Mongolia and the Nganasan in northern Siberia (Figure E.8). **C4b1** (dates to ~6 ka) is seen, albeit at lower levels, in Southern Siberia, Russia (Starikovskaya *et al.*, 2005), Eskimo Udegei in eastern Siberia (Mishmar *et al.*, 2003), Kyrgyzstan (Ingman *et al.*, 2000) and Inner Mongolia, China (Kong *et al.*, 2006).

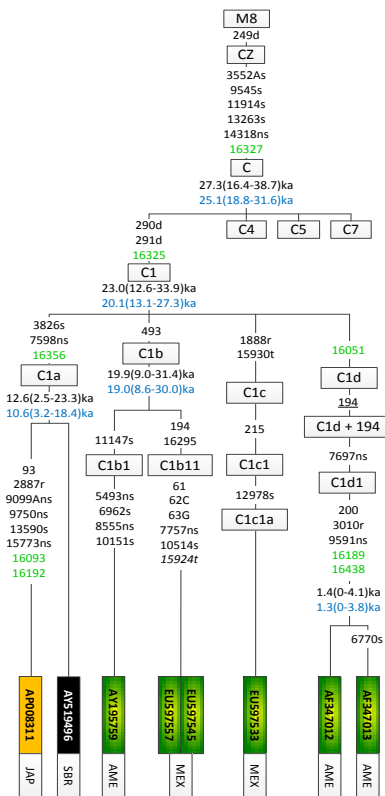


Figure E.7 The tree of haplogroup C1. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (AME – America, JAP – Japan, MEX – Mexico, SBR – Siberian Russia)

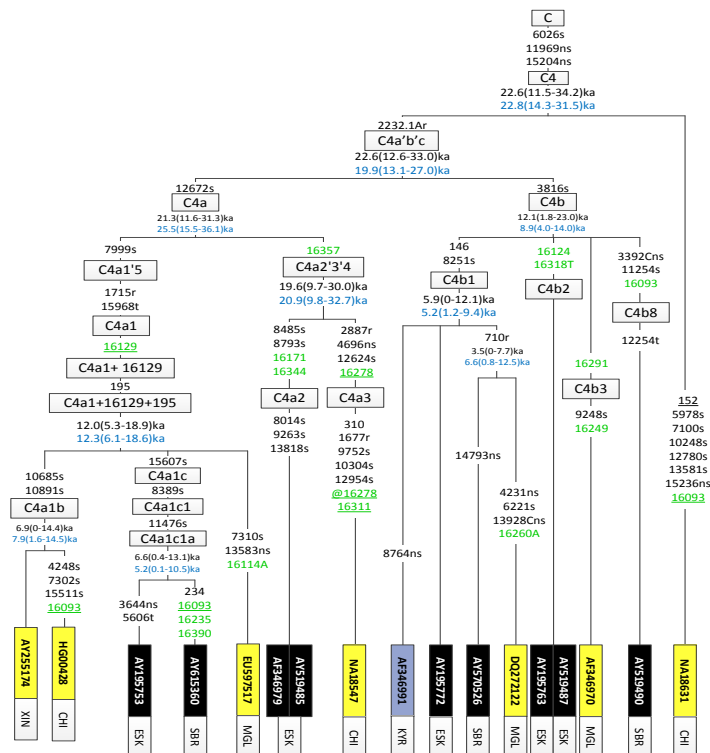


Figure E.8 The tree of haplogroup C4. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, ESK – Eskimo, KYR – Kyrgyzstan, MGL – Inner Mongolia, China, SBR – Siberian Russia, XIN – Xinjiang, China)

## Haplogroup M10

**M10** dates to ~33 ka, and can be divided into the major M10a (Yao *et al.*, 2002a; Kong *et al.*, 2003a) and the minor M10b; the latter is a newly proposed subclade in this study (Figure E.9). M10 is reconstructed by 15 complete sequences: 13 M10a and 2 M10b. M10 is mainly found in China and Japan, and at lower levels in Vietnam and India.

**M10a** dates to ~21 ka with a basal lineage is seen in Japan (Tanaka *et al.*, 2004), as well as two Late Glacial subclades, M10a1 (~18 ka) and M10a2 (~15 ka). **M10a1** has a basal lineage in northern China (Zheng *et al.*, 2011), and a nested subclade **M10a1a**, dating to ~14 ka. **M10a1a1** dates to ~5 ka and is found in South China (Kong *et al.*, 2003a) and Japan (Tanaka *et al.*, 2004), while **M10a1a2**, dates to ~10 ka, and single instances are seen in northern China (Zheng *et al.*, 2011) and Vietnam (Archaeogenetics Research Group, Huddersfield), suggesting a migration south from China within the last 10 ka. **M10a2** is seen in North and South China (Kong *et al.*, 2003a; Zheng *et al.*, 2011) with a single individual in India (Chandrasekar *et al.*, 2009), suggesting a migration east from China since 15 ka, while **M10a2a** is restricted to North China (Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004) and dates to ~8 ka.

**M10b** has undergone very heavy drift resulting in a younger, Neolithic date of ~5 ka when compared with M10a, which dates to the LGM. M10b is found in an instance each from South China (Kong *et al.*, 2006) and Vietnam (Archaeogenetics Research Group, Huddersfield). Considering the overall whole-mtDNA tree, M10 appears to have a northern origin in China, and may have arrived in Vietnam since the early Holocene (considering both M10a1a2 and M10b, and assuming they might have dispersed together).

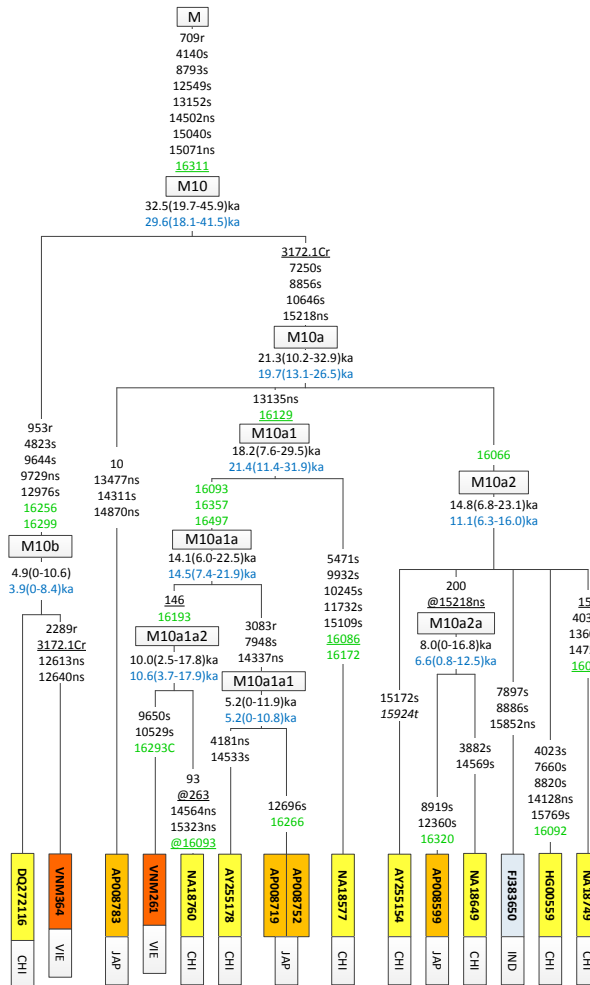


Figure E.9 The tree of haplogroup M10. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, IND – India, JAP – Japan, VIE – Vietnam)

## Haplogroup M60

**M60** is basal to haplogroup M, and dates to ~36 ka (Figure E.10). It is divided into M60a (~23 ka) and M60b (~33 ka), and both are mainly restricted to northeast India (Chandrasekar *et al.*, 2009). M60b is found elsewhere in Palangkaraya of South Borneo (Archaeogenetics Research Group, Huddersfield). Only one possible M60 lineage is found in the HVS-I database in Banjarmasin, South Borneo, which shares the M60b HVS-I motif 16266 and 16284 with the Palangkaraya sample, with a further transition at np 16290. However, since there is limited number of complete sequences, it is likely a recent migration from India to Borneo.

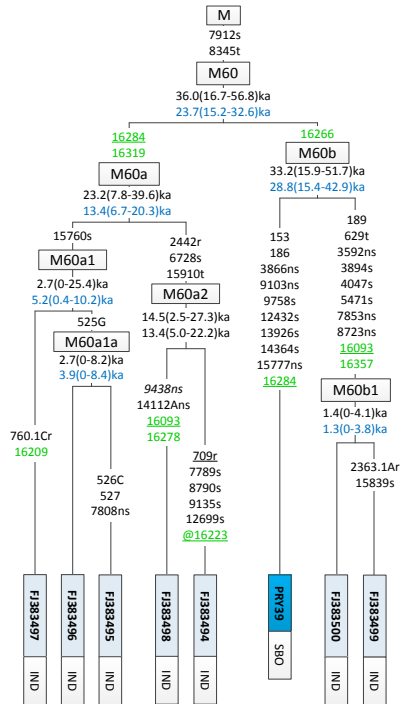


Figure E.10 The tree of haplogroup M60. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (IND – India, SBO – South Borneo)

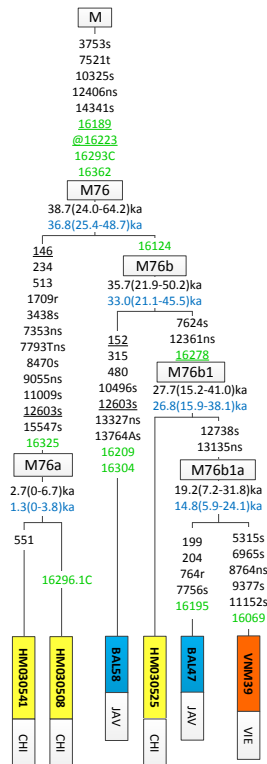


Figure E.11 The tree of haplogroup M76. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAV – Java, Indonesia, VIE – Vietnam)

## Haplogroup M76

**M76** dates to ~39 ka and it is divided into M76a and M76b (Figure E.11). Although extremely rare, M76 is restricted to South China/Sunda and may have originated in the region. **M76a**, dates recently to ~3 ka, is seen in two South Chinese individuals (Kong *et al.*, 2011). **M76b** and M76b1 pre-date the LGM; a basal lineage is seen in Java (Archaeogenetics Research Group, Huddersfield), a nested lineage in South China (Kong *et al.*, 2011), and a further nested subclade shared between Vietnam and Java (Archaeogenetics Research Group, Huddersfield), dating to the end of the LGM, ~19 ka.

## Haplogroup D4

In Figure E.12, haplogroup **D1** dates to the LGM ~22 ka, and it is seen in Native Americans (Ingman *et al.*, 2000; Mishmar *et al.*, 2003) and Brazil (Hartmann *et al.*, 2009). **D4b** dates to ~28 ka, and as mentioned earlier, D4b is mainly seen in Japan and at lower levels in China. D4b1 dates to ~22 ka, and nested within are D4b1a and D4b1b'd. **D4b1a** splits into D4b1a1 and D4b1a2a, the latter is not dated due to time constraint, which is seen in northern China (Zheng *et al.*, 2011). **D4b1a1** dates to ~3 ka, and its subclade **D4b1a1a** ~2 ka, found only in Japan (Tanaka *et al.*, 2004).



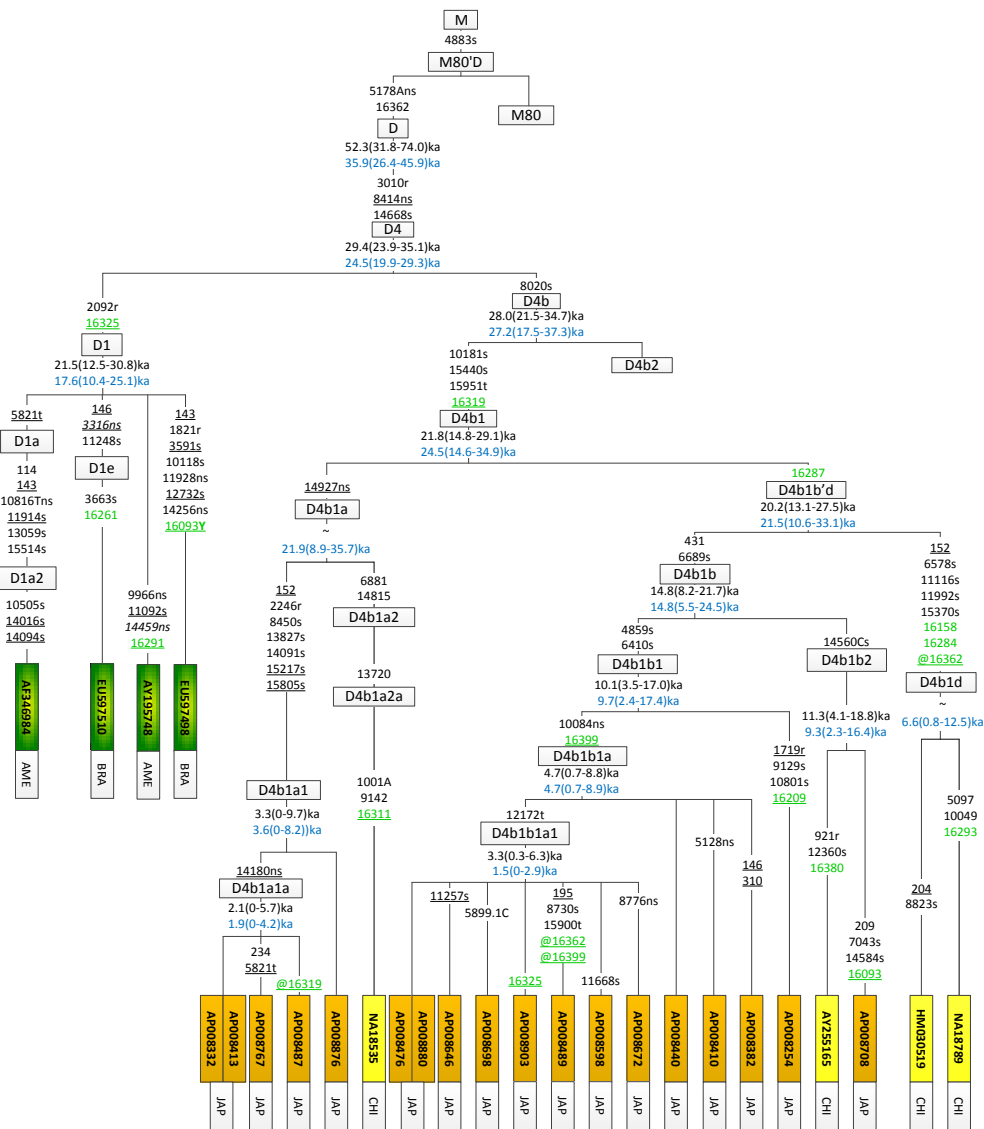
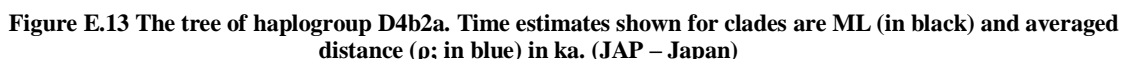


Figure E.12 The tree of haplogroups D1 and D4b1. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AME – America, BRA – Brazil, CHI – China, JAP – Japan)



293

expansion was from China to Japan, happened perhaps during the Late Glacial/early Holocene before Japan split off from the mainland.

**D4b2** also dates to the LGM ~20 ka, and has two subclades: D4b2a and D4b2b. As shown in Figure E.13, **D4b2a** dates to ~18 ka and its subclades **D4b2a1** and **D4b2a2** are only found in Japan (Tanaka *et al.*, 2004). Again, suggesting that Northeast Asia was settled early by D4 carriers in the Late Glacial. In Figure E.14, **D4b2b** dates to ~13 ka, and includes subclades D4b2b1, D4b2b2, D4b2b3, a rare subclade D4b2b4 with four paraphyletic lineages (one Japan, two northern China and one South China). The root type of D4b2b is also found in the Laotian HVS-I data (Bodner *et al.*, 2011), which implies dispersals of D4b2b south into North MSEA. **D4b2b1** dates to ~7 ka, and is found only in Japan (Tanaka *et al.*, 2004). **D4b2b2** dates to ~11 ka, and is found in northern China (Kong *et al.*, 2011; Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004). **D4b2b3** dates to ~4 ka and found in Japan (Tanaka *et al.*, 2004). As a whole, the whole-mtDNA provides evidence that the clade D4b2b overall spread from south to north as indicated by the basal lineages in South China and the HVS-I root type in NMSEA since the main northern subclades being mid to late Holocene.

**D4c** dates to ~25 ka, and is subdivided into D4c1 and D4c2 (Figure E.15). It is entirely a Japanese (Tanaka *et al.*, 2004) haplogroup. **D4c1**, dating to ~20 ka, and includes **D4c1a** (~6 ka) and **D4c1b** (~10 ka). On the other hand, a young subclade nested within D4c, **D4c2c**, dating to ~3 ka and confined to Japan. Although the root types of D4c are found in South China in the HVS-I database (Archaeogenetics Research Group, Huddersfield), it still looks to have a very early northern expansion.

**D4e** dates to ~24 ka, and includes subclades D4e1'3, D4e2 and D4e5 (Figure E.16). The root type of D4e1'3 is found in the Laotian HVS-I data (Bodner *et al.*, 2011). **D4e1** dates to ~21 ka, and it has two subclades D2 and D4e1a. D2 is subdivided into D2a and it is seen in Chukchi, Aleut and Athapaskans of northern Eskimos, while D2b found only in Siberia (Tamm *et al.*, 2007). In the whole-mtDNA tree, D2, or rather **D2a1b** is represented here by a singleton from Siberia, Russia (Ingman *et al.*, 2000).



are found in Japan (Tanaka *et al.*, 2004), both date to ~4 ka and ~2 ka respectively. The Thai lineage can therefore be inferred as intrusive dispersal from the north since there is just the one and there are quite a few Japanese lineages.

**D4e2** dates to ~5 ka, and subclades D4e2a, D4e2b, D4e2c and an unnamed subclade defined by a transition at np 16129, all date between ~2 ka and ~4 ka. **D4e5**, on the other hand, is represented by a single instance. D4e is entirely seen in Japan, and its starlike phylogeny suggests a population expansion there ~5 ka (Tanaka *et al.*, 2004).

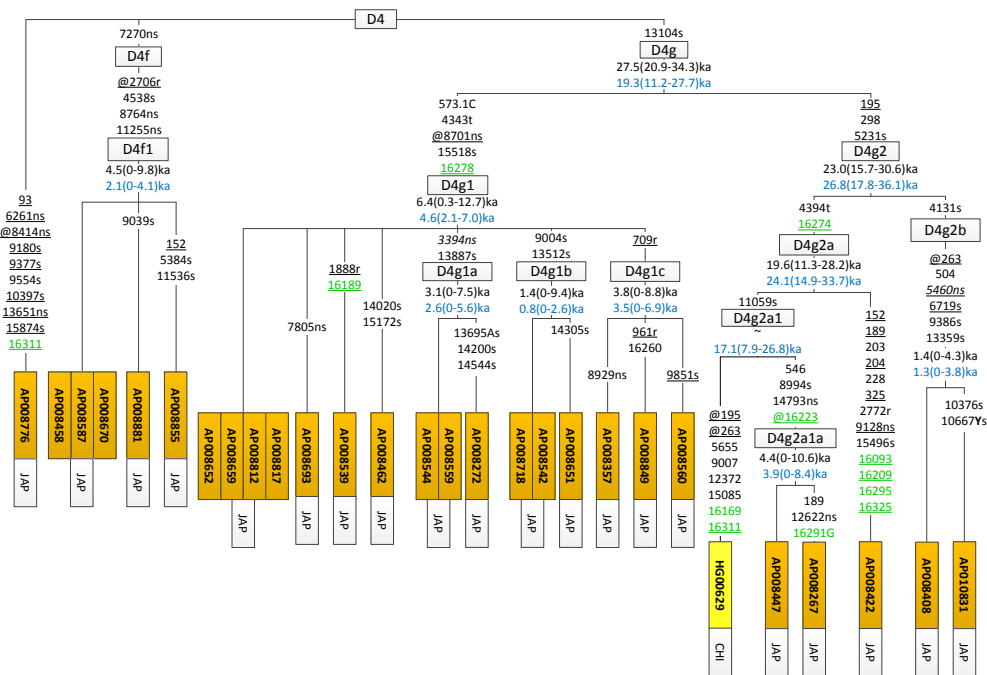


Figure E.17 The tree of haplogroups **D4f** and **D4g**. Time estimates shown for clades are ML (in black) and averaged distance ( $p$ ; in blue) in ka. (CHI – China, JAP – Japan)

Figure E.17 shows haplogroup **D4f**, dating to ~5 ka, and is seen only in Japan (Tanaka *et al.*, 2004). **D4g** (~28 ka) and it is divided into D4g1 and D4g2, both are largely restricted to Japan (Tanaka *et al.*, 2004). **D4g1** dates to ~6 ka, and nested within including subclades D4g1a, D4g1b and D4g1c, all date between ~1 ka and ~4 ka. **D4g2** dates to ~23 ka, which again is further subdivided into D4g2a (~20 ka) and D4g2b (~1 ka). D4g2a1 is found in South China (Zheng *et al.*, 2011) and Japan (Tanaka *et al.*, 2004). While haplogroup D4g is largely confined to Japan, the HVS-I sequences shown potentially the root type for D4g1 is found in China and D4g2a in Laos (Bodner *et al.*, 2011), China and Taiwan (Kong *et al.*, 2006; Metspalu *et al.*, 2006).

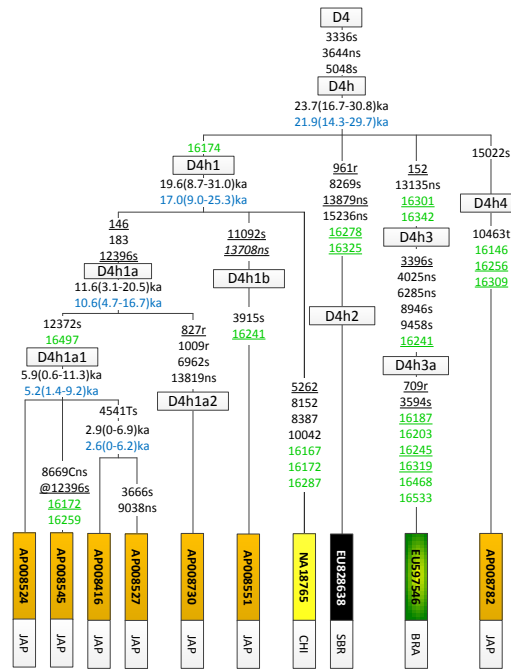


Figure E.18 The tree of haplogroup D4h. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (BRA – Brazil, CHI – China, JAP – Japan, SBR – Siberian, Russia)

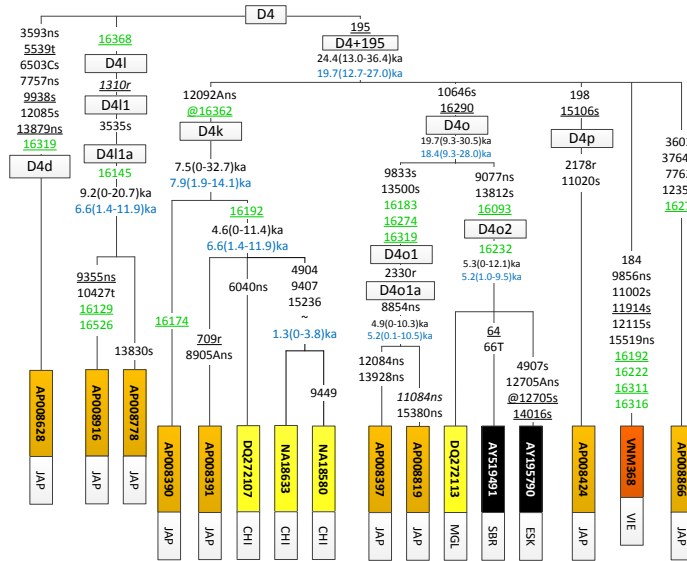


Figure E.19 The tree of haplogroups D4d, D4h1a, D4k, D4o and D4p. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, ESK – Eskimo, JAP – Japan, MGL – Inner Mongolia, China, SBR – Siberian, Russia, VIE – Vietnam)

Figure E.18 shows **D4h** dates to ~24 ka, and includes subclades D4h1, D4h2, D4h3 and D4h4. However, D4h2, D4h3 and D4h4 are represented each by an instance from Russian Far East (Starikovskaya *et al.*, 2005), Brazil (Hartmann *et al.*, 2009) and Japan (Tanaka *et al.*, 2004). Tamm *et al.* (2007) reported that haplogroup D4h3 is found between Alaska to Tierra del Fuego, and has been recently identified in Alaskan skeletal remains dating to ~10.3 ka (Kemp *et al.*, 2007). **D4h1** dates to ~20 ka and found in Japan (Tanaka *et al.*, 2004) and

North China (Zheng *et al.*, 2011). The subclades of D4h1 are all found in Japan only (Tanaka *et al.*, 2004). On the other hand, the HVS-I data shows the root type is widely distributed across East Asia albeit at a minor frequency (Kong *et al.*, 2006). The overall haplogroup D4h suggests early settlements in Japan during the LGM and a likely southern source in China.

In Figure E.19, **D4d** and **D4p** are represented here by an instance each from Japan (Tanaka *et al.*, 2004). D4l, as **D4l1a**, dating to ~9 ka and seen in two individuals from Japan (Tanaka *et al.*, 2004). Subclade **D4** with a further transition at np 195 dates to ~24 ka where the basal lineages are seen in Japan (Tanaka *et al.*, 2004) and Vietnam (Archaeogenetics Research Group, Huddersfield) and the haplogroup further subdivided into D4k, D4o and D4p. **D4k** dates to ~8 ka, and the nesting of its subclades suggest that lineages may have migrated from Japan (Tanaka *et al.*, 2004) to Qinghai and Beijing in northern China (Kong *et al.*, 2006; Zheng *et al.*, 2011) within ~5 ka (the Zheng *et al.*, 2011 sequences were not included in the ML estimations due to time constraints). However, this nesting pattern is unusual as most of the dispersals go from the southern mainland into Japan and not the other way round, therefore it may be an artefact of the few samples. **D4o** (~20 ka) includes two subclades: **D4o1a**, ~5 ka, is seen in Japan, and **D4o2**, ~5 ka, in Inner Mongolia, China (Kong *et al.*, 2006), Northern Siberia (Starikovskaya *et al.*, 2005) and Eskimo in Russian Far East (Mishmar *et al.*, 2003).

In Figure E.20, **D4i** dates to ~18 ka, and is mainly seen in Japan (Tanaka *et al.*, 2004) and North China (Zheng *et al.*, 2011). **D4j** dates to ~17 ka, and the paraphyletic lineages are seen in Japan (Tanaka *et al.*, 2004), South and North China (Kong *et al.*, 2003a; Zheng *et al.*, 2011). Subclade D4j3a is seen in Japan (Tanaka *et al.*, 2004) and Beijing, China (Zheng *et al.*, 2011), but I was not able to date it due to time constraints, same for D4m. **D4m** is seen in a sequence from North China (Zheng *et al.*, 2011), and nested within is a subclade, **D4m1** ~4 ka, seen in Japan (Tanaka *et al.*, 2004). **D4n** dates to ~8 ka, and subclades D4n1 and D4n1a both date to ~3 ka and ~1 ka respectively, found entirely in Japan (Tanaka *et al.*, 2004). Consistent with the majority of D4 carriers, subclades D4i, D4j, D4m and D4n are likely to have a Late Glacial source in China, potentially somewhere in the southern region.

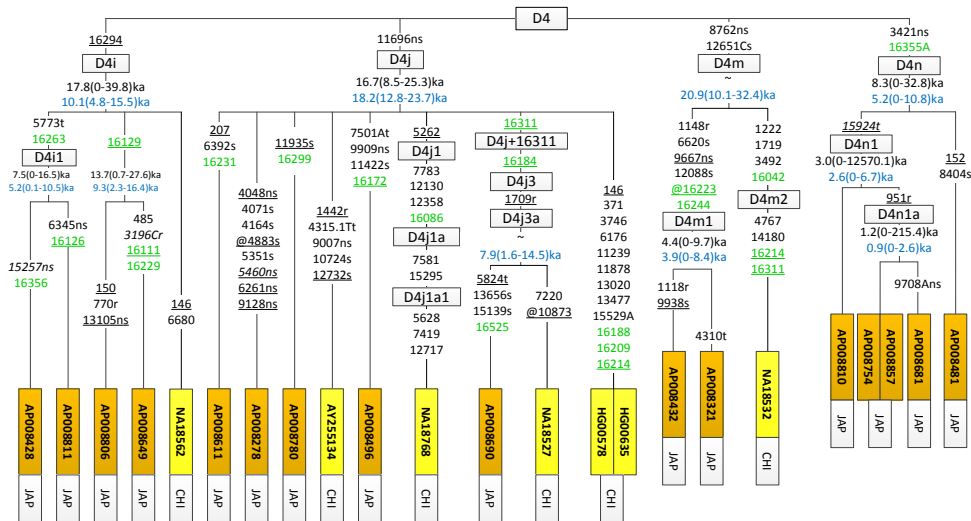


Figure E.20 The tree of haplogroups D4i, D4j, D4m and D4n. Time estimates shown for clades are ML (in black) and averaged distance (p; in blue) in ka. (CHI – China, JAP – Japan)

## Haplogroup D5a

In Figure E.21, **D5a** dates to ~29 ka and it includes subclades D5a1, D5a2 and D5a3. **D5a1** is seen in Japan (Tanaka *et al.*, 2004) with an estimated age of ~9 ka. **D5a3** is represented by an instance from Russian Siberia (Starikovskaya *et al.*, 2005). The branch of D5a with a reversion at np 1438 consists of subclade D5a2, as **D5a2a**, and is seen in a single basal branch in Japan (Tanaka *et al.*, 2004). Unfortunately, no D5a2\* lineages are represented in the tree due to time constraints, although it is found in North East India (Chandrasekar *et al.*, 2009; van Oven and Kayser, 2009). Besides, the mtDNA HVS-I database shows that D5a2 is seen in South China, indicating that that is the source, and that D5a2a is also seen much more commonly in South China than Japan and Korea in HVS-I.

D5a2 is seen in South China in HVS-I, a source in South China, and it is also seen much more commonly in South China than Japan and Korea in HVS-I (Archaeogenetics Research Group, Huddersfield). D5a2a dates to ~12 ka and it is divided into two subclades: D5a2a1'2 and D5a2a1 with a transition at np 16164 (Figure E.21). Subclade D5a2a1 and D5a2a2 are commonly seen in Japan (Tanaka *et al.*, 2004), China (Kong *et al.*, 2003a; Zheng *et al.*, 2011), and two instances from Russia (Starikovskaya *et al.*, 2005; Hartmann *et al.*, 2009), suggesting another mid-Holocene data for the entry to Northeast Asia. Subclade D5a2a+16164 shares the same defining transition np16164 with D5a2a1, quite possibly showing two parallel mutations; the basal lineages are seen in Japan (Tanaka *et al.*, 2004) and northern China (Zheng *et al.*, 2011) but no date is available due to time constraints.



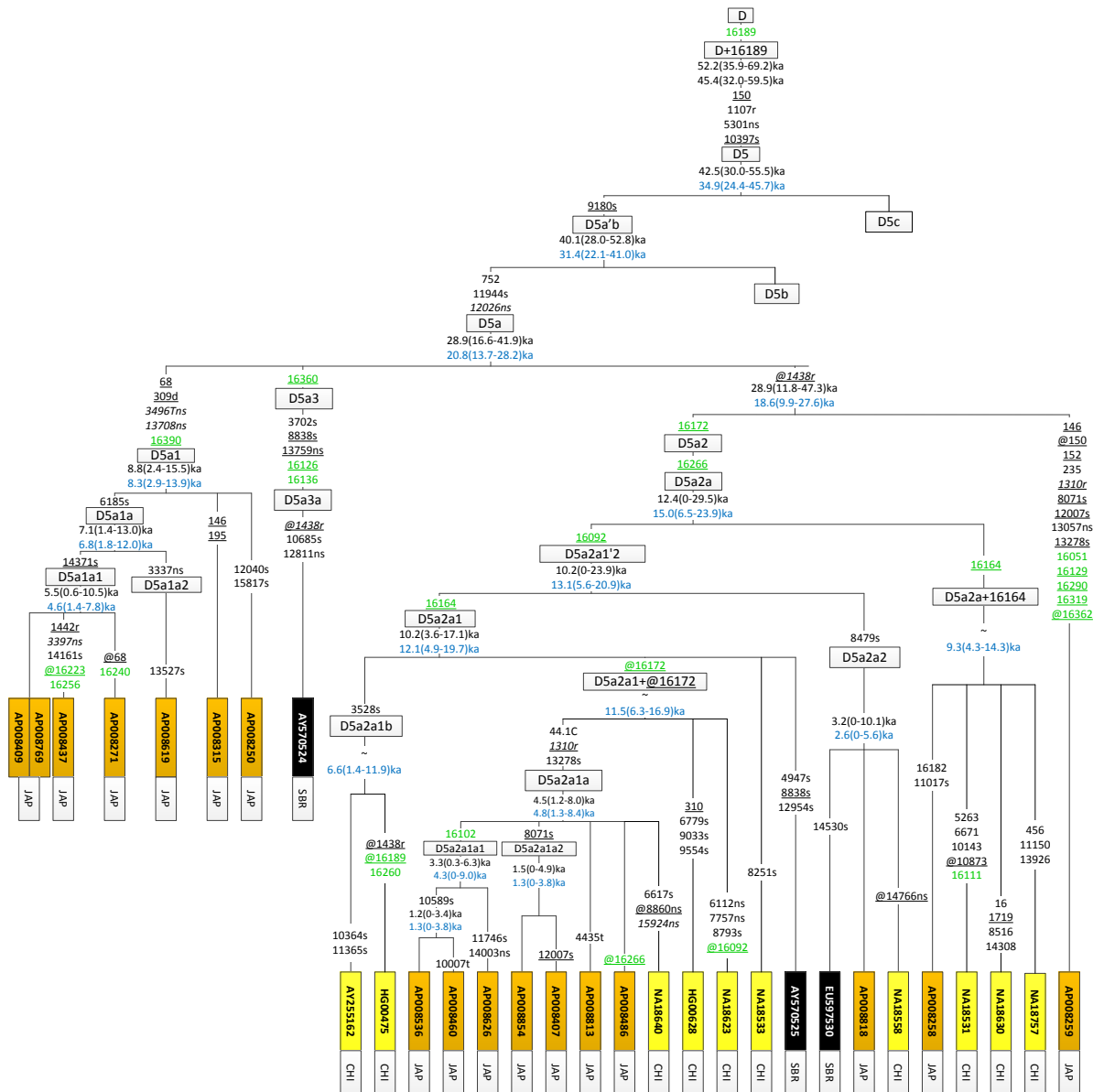


Figure E.21 The tree of haplogroup D5a. Time estimates shown for clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, SBR – Siberian Russia)

## Haplogroup N9a2'4'5

**N9a2'4'5** dates to ~18 ka and consists of N9a2, N9a4 and N9a5 (Figure E.22). **N9a2** is entirely found in Japan (Tanaka *et al.*, 2004), although an early-branching south Chinese individual reported recently by Zheng *et al.* (2011) might suggest an origin in China ~16 ka at the branch of N9a2'4'5. N9a2, dating to ~12 ka, shows a highly localised and diversified phylogeny in Japan, divided into N9a2a, N9a2c and N9a2d. **N9a2a** dates to ~6 ka. N9a2a can then be divided into N9a2a1, N9a2a2, and N9a2a3. However, only one sample is represented

in the tree for N9a2a1 (Aichi, Japan) and N9a2a3 (Tokyo, Japan; Tanaka *et al.*, 2004). N9a2a2 is detected in Chiba, Aichi and Tokyo with an age of ~2 ka.

Tanaka *et al.* (2004) reported three Japanese individuals from Aichi and Tokyo that belonged to **N9a2c**, dating to ~4 ka. Only one individual of **N9a2d** in this study is observed in Japan by Tanaka *et al.* (2004).

**N9a4** dates to ~7.5 ka, and the nesting relationships suggest an origin in Japan before spreading into China within the last few thousand years (Figure E.22). Subclade **N9a4a**, dating to ~3 ka, is only found in Aichi (Tanaka *et al.*, 2004), while **N9a4b**, at ~5 ka, has its root type found in Tokyo and spread ~2 ka into south and north China (Zheng *et al.*, 2011). **N9a5**, also a Japanese clade, dates ~8 ka. This subclade has been reported in Chiba and Tokyo by Tanaka *et al.* (2004).

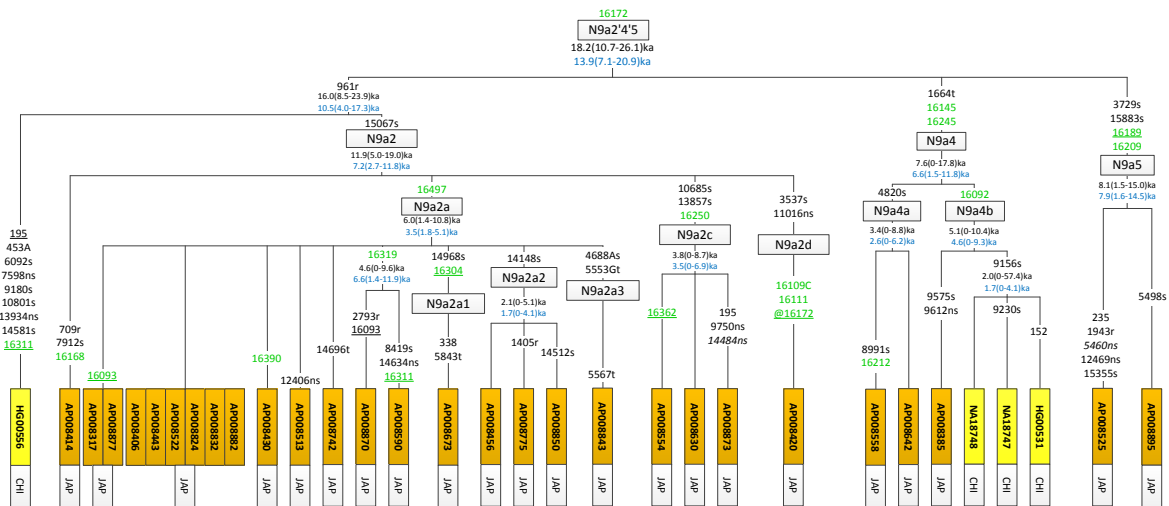
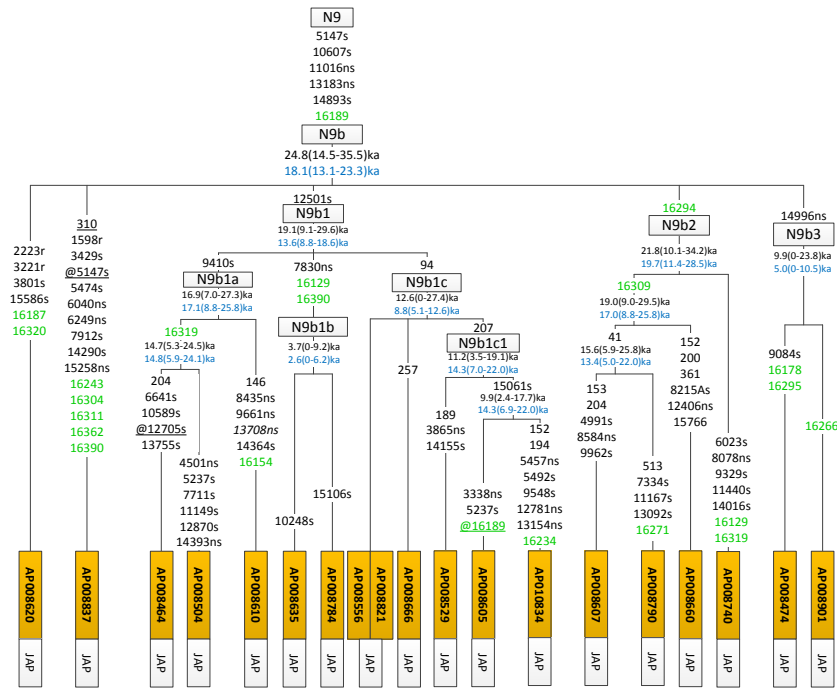


Figure E.22 The tree of haplogroup N9a2'4'5. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan)

## Haplogroup N9b

N9b can be sub-divided into N9b1, N9b2 and N9b3. Haplogroup **N9b1** diverged at ~19 ka and can be divided into N9b1a, N9b1b and N9b1c (Figure E.23). **N9b1a** has an estimated age of ~17 ka. N9b1a likely dispersed from Aichi into Chiba, Japan around ~15 ka with an additional transition at np 16319. **N9b1b** dates to ~4 ka and has been reported in two Aichi individuals. **N9b1c** dates to ~13 ka and is also found in Aichi, Japan. N9b1c1 dates to ~11 ka. Nested within is a subclade with a divergence time of ~10 ka, and is found in Aichi and Kanagawa individuals (Tanaka *et al.*, 2004).



**Figure E.23** The tree of haplogroup N9b. Time estimates shown for the clades are ML (in black) and averaged distance (p; in blue) in ka. (JAP – Japan)

**N9b2** appears to be localised in Aichi only; it is defined by a transition at np 16294 and dates to ~22 ka. A further node with an additional transition at np 16309 dates to ~19 ka, and more recently with a transition at np 41 that dates to ~16 ka. Lastly, **N9b3**, which is found in Chiba and Tokyo, dates to ~10 ka (although with p it dates to only ~5 ka).

## Haplogroup A5a

**A5a1** and its subclades are generally found in Aichi, Chiba and Tokyo (Tanaka *et al.*, 2004), with the only exception of an Inner Mongolian reported by Kong *et al.* (2006) nested within the Japanese in haplogroup A5a1a1 (Figure E.24).

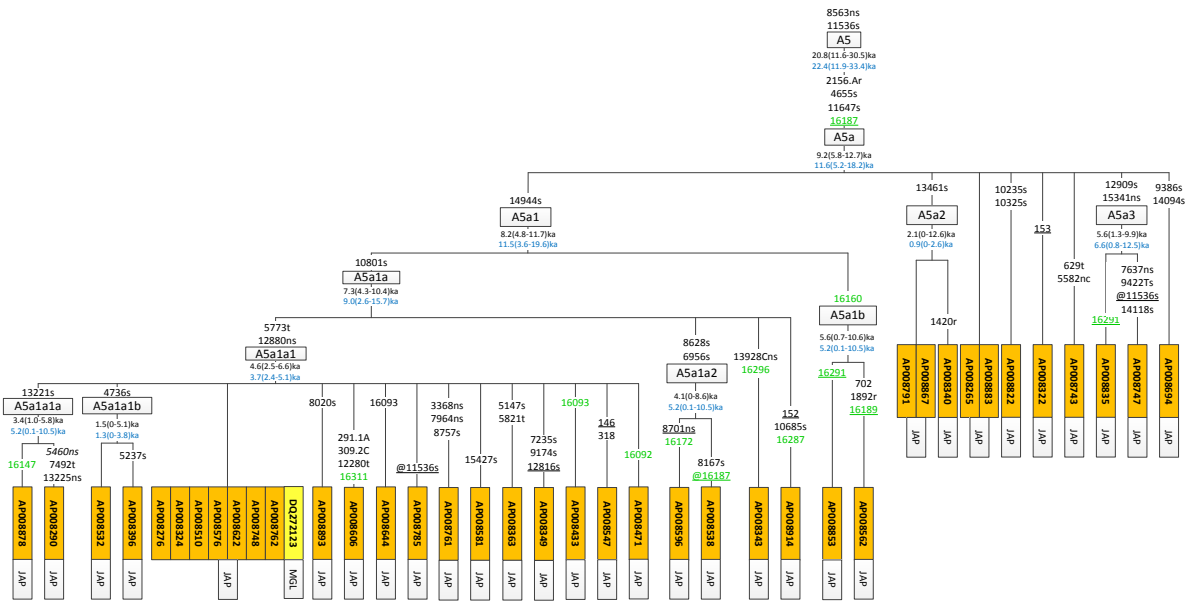


Figure E.24 The tree of haplogroup A5a. Time estimates shown for the clades are ML (in black) and averaged distance (p; in blue) in ka. (JAP – Japan, MGL – Inner Mongolia, China)

## Haplogroups B4d and B4f

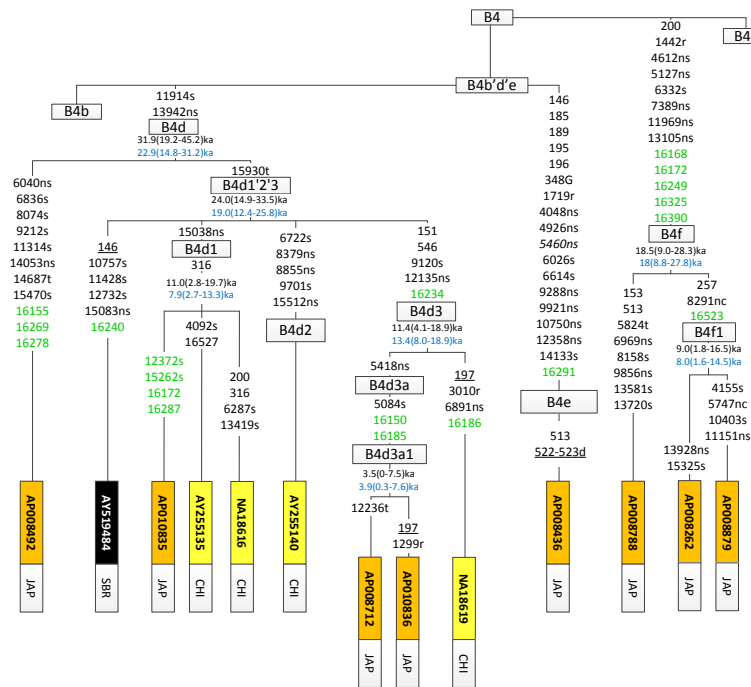


Figure E.25 The tree of haplogroup B4b'd'e and B4f. Time estimates shown for clades are ML and averaged distance (p) in ka. (CHI – China, JAP – Japan, SBR – Siberia)

**B4d** dates to ~32 ka with a basal lineage seen in Japan (Figure E.25; Tanaka *et al.*, 2004). The main subclade is **B4d1'2'3**, dating to ~24 ka, with a basal lineage seen in Siberia (Starikovskaya *et al.*, 2005). B4d1'2'3 includes subclades B4d1, B4d2 and B4d3. **B4d1** dates

to ~11 ka, and is found in China (Kong *et al.*, 2003b; Zheng *et al.*, 2011) and Japan (Nohira *et al.*, 2010). **B4d3** dates to ~11 ka, with a basal lineage in China (Zheng *et al.*, 2011), and a small derived subclade **B4d3a1**, dating to ~4 ka, in Japan (Tanaka *et al.*, 2004; Nohira *et al.*, 2010). There is just a single instance of **B4e**, seen in Japan.

**B4f** is a minor basal B4 clade dating to ~19 ka and restricted to Japan (and Korea in HVS-I) (Tanaka *et al.*, 2004). Its distribution further strengthens the phylogeographic case that B4 likely originated in China and later dispersed both into northeast Asia and SEA, as seen primarily in B4b1a2 and minor other instances.

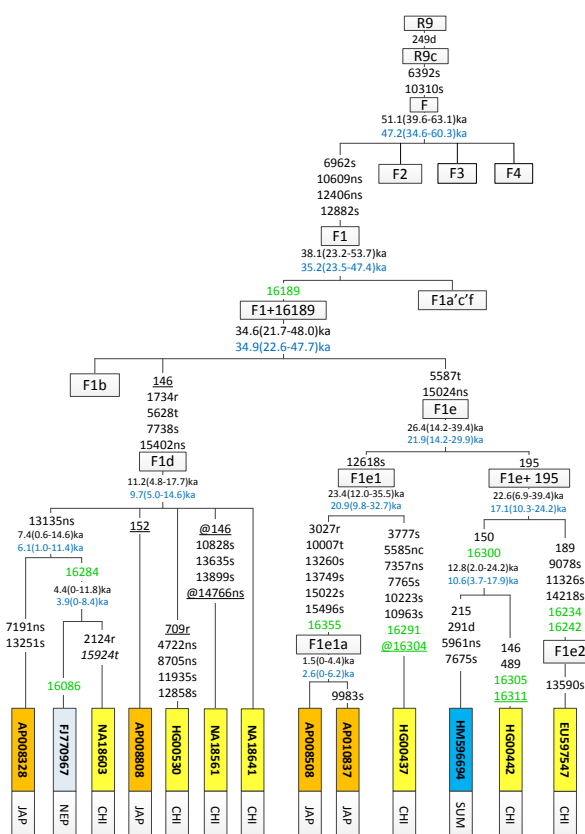
### Haplogroup B4c1

**B4c1** dates to ~35 ka, and has two main subclades, B4c1c and B4c1a'b. **B4c1c1** dates to ~3 ka and it is seen only in Japan (Tanaka *et al.*, 2004). **B4c1a'b** dates to 32 ka; it has two subclades, B4c1a and B4c1b. **B4c1a** is an East Asian clade seen mostly in Japan, with a single basal lineage in China (Tanaka *et al.*, 2004; Zheng *et al.*, 2011), and dates to ~11 ka. **B4c1b** dates to ~27 ka, and it is divided into B4c1b1 and B4c1b+16335. **B4c1b1** dates to ~6 ka and is restricted to Japan (Tanaka *et al.*, 2004). **B4c1b+16335** dates to ~24 ka and has a basal lineage in Japan, and dispersed into China ~21 ka through two branches, B4c1b2a and B4c1b2c. **B4c1b2a** is seen in China Liaoning (Kong *et al.*, 2003b) and this subclade dates to ~15 ka. **B4c1b2c** dates to ~13 ka and seen in Beijing (Zheng *et al.*, 2011). However, the HVS-I data shows that this subclade is mainly found amongst Aboriginal Taiwanese and is also widely dispersed across Southeast Asia (Archaeogenetics Research Group, Huddersfield).

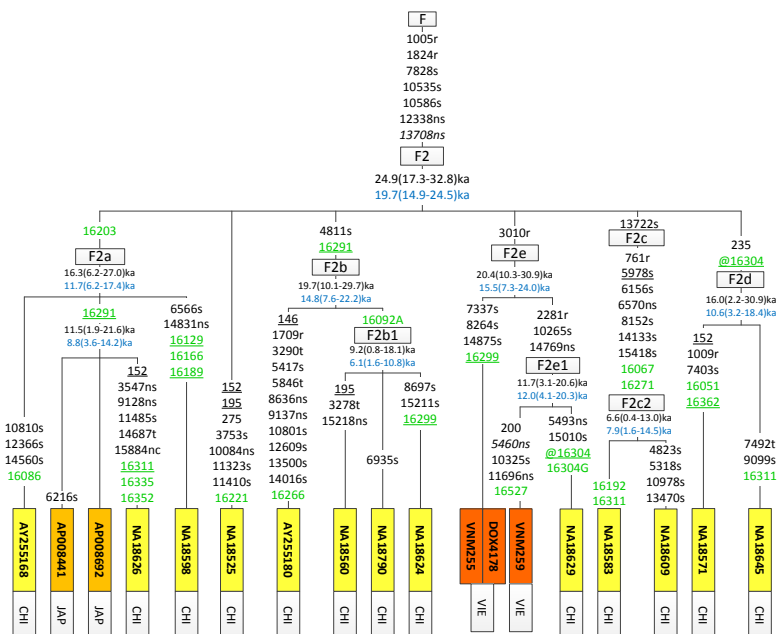
### Haplogroups F1b, F1d and F1e

In Figure E.26, **F1b1** (the sole basal branch of F1b after Kong *et al.*, 2006) dates to ~14 ka and is extremely widespread, seen in the HVS-I database across China, Northeast Asia and Central Asia, as far west as Iran, the Caucasus and Turkey. The basal lineages of F1b1 are reported from northern China (Zheng *et al.*, 2011), Japan (Tanaka *et al.*, 2004) and Siberia (Kong *et al.*, 2006), with F1b1a (confined to Japan) being the only substantial branch. F1b1a diversified locally in Japan, with starlike subclades appearing ~5–6 ka. F1b appears likely to have a northern, Late Glacial origin.





**Figure E.27** The tree of haplogroup F1d and F1e. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, NEP – Nepal, SUM – Sumatra)



**Figure E.28** The tree of haplogroup F2. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (CHI – China, JAP – Japan, VIE – Vietnam)

## Haplogroup F2

Haplogroup F2 clade dates to ~25 ka (Figure E.28). It is divided into F2a, F2b, F2c, F2d and F2e. **F2a** dates to ~16 ka, with the basal types seen in northern China (Kong *et al.*, 2003b; Zheng *et al.*, 2011). A subclade defined by a transition at np 16291 and estimated at ~12 ka is found in China and Japan (Tanaka *et al.*, 2004; Zheng *et al.*, 2011) and, much more rarely, in Vietnam and aboriginal Taiwanese (Mormina, 2007). **F2b** dates to ~20 ka and F2b1 ~9 ka is reported in northern China (Kong *et al.*, 2003b; Zheng *et al.*, 2011) with occasional examples in Thailand (Mormina, 2007). **F2c2**, dating to ~7 ka, and **F2d**, dating to ~16 ka, are both seen in northern China (Zheng *et al.*, 2011). **F2e** dates to ~20 ka, including **F2e1**, dating to ~12 ka, is reported from Vietnam (Archaeogenetics Research Group, Huddersfield) and China (Zheng *et al.*, 2011). Overall, F2 is a haplogroup broadly centred on China that point to minor instances of gene flow both north and south.

## Haplogroup P

**P2'10** dates to ~63 and it is divided into P2 and P10. Haplogroups **P2** dates to ~30 ka, where the whole clade is restricted to PNG (Ingman and Gyllensten, 2003; Pierson *et al.*, 2006; Friedlaender *et al.*, 2007). **P10** undergone high drift and dates to Neolithic ~5 ka where it is found in the Philippines (Tabbada *et al.*, 2010 and Archaeogenetics Research Group, Huddersfield).

**P3** dates to ~49 ka, and is divided into P3a and P3b (Figure E.29). **P3a** dates to ~30 ka and is found in the Australian Aboriginals (Ingman and Gyllensten, 2003). **P3b** dates to ~42 ka in PNG, and its subclade P3b1 is seen in Australia (Friedlaender *et al.*, 2007) dating to the LGM ~25 ka.

**P4** dates to ~63 ka, and it is divided into P4a and P4b (Figure E.29). **P4a** dates to ~23 ka and together with its subclade P4a1 are predominantly found in PNG (Friedlaender *et al.*, 2007), with a basal lineage seen in Sulawesi, Indonesia (Archaeogenetics Research Group, Huddersfield). **P4b** (~48 ka) and P6 (~62 ka), similar to P3a, are found in the Australian Aboriginals (Ingman and Gyllensten, 2003). Besides, three basal to haplogroup P lineages: two from the Australian Aboriginals was proposed by Ingman and Gyllensten (2003) as 'subclades' **P5** and **P7**, and one Filipino as **P9** (Tabbada *et al.*, 2010).



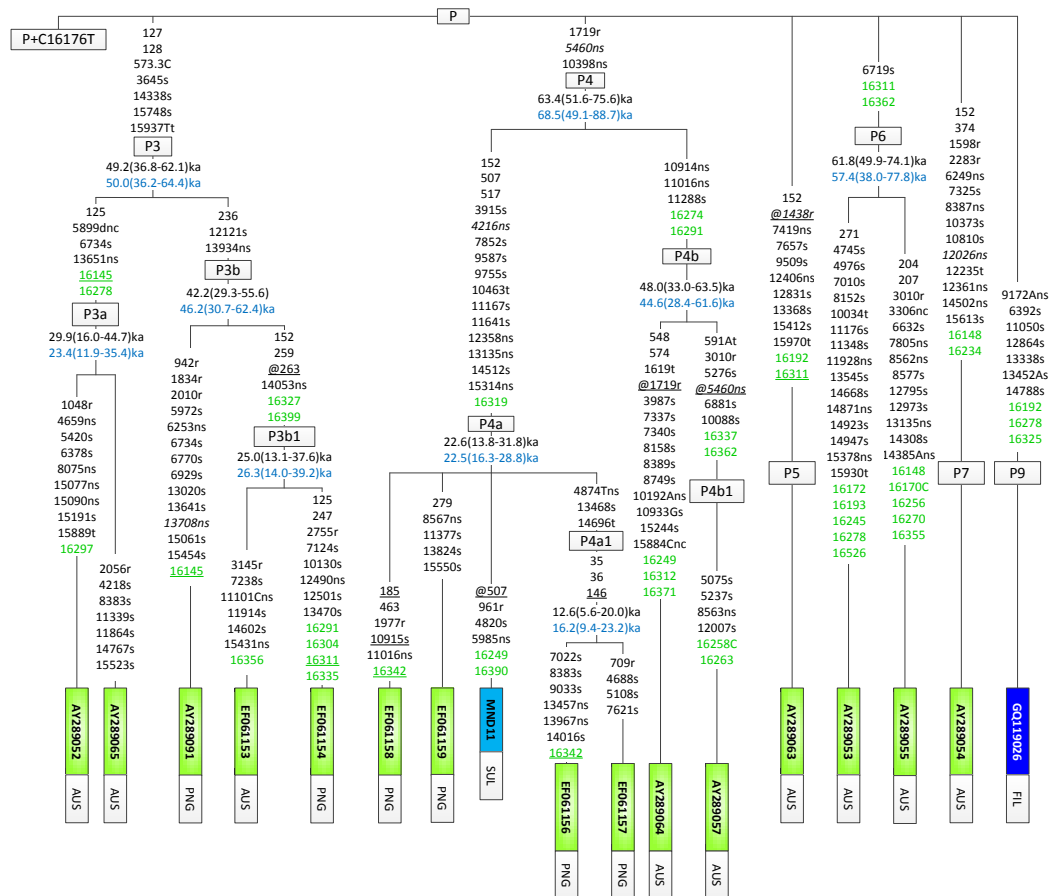


Figure E.29 The tree of haplogroups P3, P4, P5, P6, P7 and P9. Time estimates shown for the clades are ML (in black) and averaged distance ( $\rho$ ; in blue) in ka. (AUS – Australia, FIL – Philippines, SUL – Sulawesi, PNG – Papua New Guinea)

## Haplogroup R30

The subclades of this rare R30 haplogroup have undergone genetic drift. **R30a** dates to ~ 37 ka (Figure E.30) and a single basal lineage is seen in Sumatra (Archaeogenetics Research Group, Huddersfield), while nested within is a smaller subclade seen in Sri Lanka (Chaubey *et al.*, 2008) and Nepal (Fornarino *et al.*, 2009). A single instance representing R30b (**R30b1**) is found in Punjab of northern India (Chaubey *et al.*, 2008). Generally, R30 shows early expansions from India into SEA and the relict descendant is found in Sumatra.

